Drug repurposing for the treatment of COVID-19: a knowledge graph approach

Vincent KC Yan[†], Xiaodong Li[†], Xuxiao Ye[†], Min Ou, Ruibang Luo, Qingpeng Zhang, Bo Tang, Benjamin J Cowling, Ivan Hung, Chung Wah Siu, Ian CK Wong, Reynold CK Cheng[*], Esther W Chan[*]

[†] Vincent KC Yan, Xiaodong Li and Xuxiao Ye contributed equally to this work. [*] Esther W Chan and Reynold CK Cheng share senior authorship.

V. K. Yan, X. Ye, Prof. I. C. Wong, Dr. E. W. Chan
Centre for Safe Medication Practice and Research
Department of Pharmacology and Pharmacy
LKS Faculty of Medicine
University of Hong Kong
Hong Kong Special Administrative Region, China
1/F, Jockey Club Building for Interdisciplinary Research, 5 Sassoon Road, Pokfulam, Hong Kong SAR, China
Email: ewchan@hku.hk

Dr. X. Li, M. Ou, Dr. R. Luo, Prof. R. C. Cheng
Department of Computer Science
Faculty of Engineering
University of Hong Kong
Hong Kong Special Administrative Region, China
CB303, Chow Yei Ching Building, Pokfula, Hong Kong SAR, China
Email: ckcheng@cs.hku.hk

Dr. Q. Zhang
School of Data Science
City University of Hong Kong
Hong Kong Special Administrative Region, China
83 Tat Chee Avenue, Kowloon, Hong Kong SAR, China

Dr. B. Tang
Department of Computer Science and Engineering
Southern University of Science and Technology
China
1088 Xueyuan Avenue, Nanshan District, Shenzhen, Guangdong, China

Prof. B. J. Cowling
Division of Epidemiology and Biostatistics
School of Public Health
University of Hong Kong
Hong Kong Special Administrative Region, China
21 Sassoon Road, Pokfulam, Hong Kong SAR, China

Prof I. Hung
Division of Infectious Diseases
Department of Medicine

LKS Faculty of Medicine
University of Hong Kong
Hong Kong Special Administrative Region, China
102 Pokfulam Road, Hong Kong SAR, China

Prof C. Siu
Division of Cardiology
Department of Medicine
University of Hong Kong
Hong Kong Special Administrative Region, China
102 Pokfulam Road, Hong Kong SAR, China

Keywords: COVID-19; drug repurposing; knowledge graph; ranking; motif scores

Abstract

Background

Identifying effective drug treatments for COVID-19 is essential to reduce morbidity and mortality. Although a number of existing drugs have been proposed as potential COVID-19 treatments, effective data platforms and algorithms to prioritise drug candidates for evaluation and application of knowledge graph for drug repurposing have not been adequately explored.

Methods

A COVID-19 knowledge graph is developed by integrating 14 public bioinformatic databases containing information on drugs, genes, proteins, viruses, diseases, symptoms and their linkages. An algorithm is developed to extract hidden linkages connecting drugs and COVID-19 from the knowledge graph, to generate and rank proposed drug candidates for repurposing as treatments for COVID-19 by integrating three scores for each drug: motif scores, knowledge graph PageRank scores and knowledge graph embedding scores.

Results

A knowledge graph containing over 48,000 nodes and 1,337,000 edges is developed, which includes 13,563 molecules in the DrugBank database. From the 5,624 molecules identified by the motif-discovery algorithms, ranking results show that 112 drug molecules have the top 2% scores, of which 50 existing drugs with other indications approved by health administrations are reported. These include 12 cardiovascular agents (8 statins, moexipril, isosorbide mononitrate/dinitrate, spironolactone, bezafibrate), 11 anti-infectives (4 antiviral agents, 2 antiparasitic agents, 3 antibacterial agents, 2 antifungal agents), 10 hypnotics or anticonvulsants, 7 antineoplastic agents, 1 immunosuppressant, 4 hormonal agents and 4 other agents (pirfenidone, ibuprofen, amitriptyline, dexamethasone, fostamatinib) which are promising, with some worthy of further investigation.

Conclusion

A novel knowledge graph approach can identify potential drug candidates for repurposing as treatment of COVID-19. The proposed drug candidates serve to generate hypotheses for future evaluation in clinical trials and observational studies.

1. Introduction

COVID-19 has emerged as a severe pandemic with a high transmission rate and significant mortality. By the middle of April 2021, there were over 140 million confirmed cases globally.[1] The lack of specific drug treatment for COVID-19 has contributed to more than 3 million deaths worldwide.[1] To date, two mRNA and one adenoviral vector COVID-19 vaccines were granted emergency use authorization (EUA) in the United States and development of COVID-19 vaccines in other countries is ongoing.[2-4] However, the safety of COVID-19 vaccines in general remains a concern as multiple serious adverse events such as Bell's palsy and thrombosis had been reported with their use.[5] Questions also remain about the efficacy of COVID-19 vaccines since the duration of protection, efficacy in populations excluded from the trials, and robustness against mutations of SARS-CoV-2 have not been evaluated. As such, discovering effective drug treatment for COVID-19 remains essential.

Over the past year, a number of drug candidates, mainly antiviral agents and monoclonal antibodies, were evaluated for their efficacy as COVID-19 treatments. Yet, preliminary results suggest that some of these agents may not be as promising as speculated. For instance, although the United States Food and Drug Administration approved remdesivir for hospitalised patients with COVID-19 aged 12 years old or above regardless of disease severity, the optimal role and benefit of remdesivir remain controversial since there is no clear evidence of mortality reduction in clinical trials, leading to recommendations of not using it by the World Health Organisation.[6] Other drug candidates are associated with serious adverse effects, such as electrocardiographic changes with hydroxychloroquine, which limits their usage.[7,8] Hence, effective data platforms and tools are essential to enable efficient identification of new drug candidates in search of safer and more efficacious alternatives. While conventional structure-based screening methods such as protein docking analyses are traditionally used for *de novo* drug discovery, repurposing existing drugs provides a more cost and time efficient means of discovering treatment for new diseases.[9-11] Various approaches, including network-based, structure-based and AI-based approaches for drug repositioning had been investigated, yet the application of knowledge graph in this domain warrants further exploration.[12]

Previous studies applied knowledge graphs to different research domains in medicine, including disease subtyping[13] and herb recommendation.[14] Current studies on COVID-19 knowledge graphs are largely based on literature mining,[15] and linking COVID-19 publications, case statistics and genes.[16] However, these knowledge graphs are often limited in scale and while some may include drug-target information, no single knowledge graph is fully unified with integrated information for drug discovery, including drug-protein, drug-gene relationships and protein domain information which provides an essential bridge between genes, proteins, drugs, viruses and diseases.[17] Also, efficient algorithms providing a ranking of drug candidates utilizing information from large-scale knowledge graphs have not been explored.

In this study, we applied a knowledge graph-based method to identify potential drug candidates for repurposing as COVID-19 treatment. The knowledge graph integrates known relations between viruses (including SARS-CoV-2), drugs, genes, proteins, diseases, symptoms and from multiple large-scale open data sources. The results will generate hypotheses of potential drug candidates which can be further tested via clinical trials and observational studies. Healthcare professionals and other researchers will also be able to tune the algorithms (for example, give more weight on specific edges such as symptoms) to generate personalised drug ranking results.

2. Methods

2.1. Building a COVID-19 knowledge graph for drug repurposing

Knowledge graphs enable identification of valuable information regarding the large-scale, complex relationships among different entities. Knowledge graph is a multi-relational graph composed of entities (nodes) and relations (edges).[18] In the case of a COVID-19 knowledge graph for drug repurposing, each node represents a specific protein, gene, drug, virus, disease or symptom, whereas each edge represents a known existing linkage between any two nodes (**Figure 1**). Data on linkages from different data sources were processed into the corresponding nodes (see “Data sources”) and edges (**Table 1**), thus integrating known relations from disparate data sources into a large-scale knowledge graph. Drug repurposing algorithms were then used to extract hidden linkages about drugs and COVID-19 from the knowledge graph, and further ranked using computational scoring methods, to shortlist potential drug candidates for COVID-19 drug repurposing. It should be noted that no explicit linkages between any drug and SARS-CoV-2 is present in the knowledge graph, since high-quality evidence on effective treatments for COVID-19 remains scant,[6] and main aim of this study is to develop a method to propose drug candidates in the absence of data on definite drug-virus relationships.

2.2. Data sources

We collected data from large-scale open data sources in three broad bioinformatic categories: drug-target interactions, gene-gene interactome, and gene-disease network. Data on drug-target interactions comprised drug metadata and drug-target linkages. Drug metadata were retrieved from DrugBank with relevant clinical trials information from ClinicalTrials.gov. Drug-target linkages were collected from the Pharmacogenomics Knowledgebase (PharmGKB), BindingDB, Therapeutic Target Database, and DrugBank, and were further filtered by binding affinities and review status from UniProt. Data on gene-gene interactome were collected from BioGRID, Database of Interacting Proteins, and Human Protein Reference Database. Data on gene-disease network were collected from Comparative Toxicogenomic Database, and Human Phenotype Ontology (HPO) database. Further details of the data sources and data integration process are described in Supporting Information.

2.3. Data pre-processing and integration

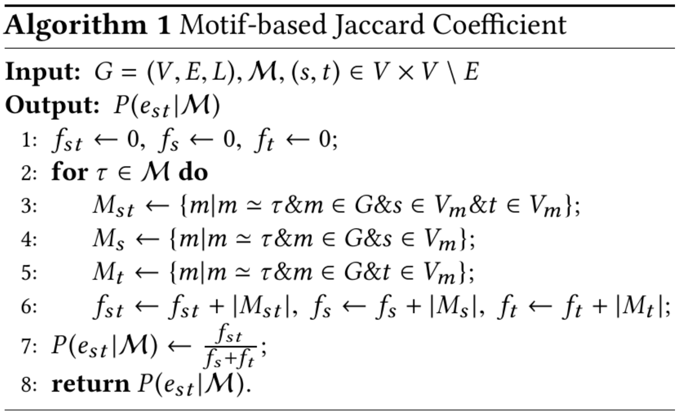
The DrugBank ID was used to represent each drug in the graph. The NCBI Entrez ID and official gene symbol were used to represent the gene while the mapping information of the gene and protein was retrieved from UniProt.[19] Disease mapping was based on the Disease Ontology database,[20] while Medical Subject Headings (MeSH) ID was used to represent each disease in the graph.[21] To align the data from different sources, records from terminology databases such as HPO which provides unique identifiers for entities with different alias were used. Databases that consist of genes, proteins, diseases, drugs, and pathways, were integrated into the knowledge graph by the publicly-used IDs in order to support information retrieval and further cross-validation. For databases with genes (the drug-target interactions, gene-gene interactome, and Gene-disease network), the NCBI Gene ID was used as the unified ID for record import. Since biological databases might also use the name of the protein product to represent the gene, the UniProt ID and the official gene symbol from NCBI were used to match the protein records to the gene records. For databases that involve drugs and drug-target interactions, each of them has a set of in-house drug IDs, but the drug name or its synonyms are standardised. These databases were merged based on drug names and the mapping was verified by pharmacists. For databases that provide linkages of gene and diseases (Gene-disease network), the Disease Ontology was used, which provided commonly-used disease ID mappings that was used to convert other disease IDs into MeSH IDs.

2.4. Extraction and ranking of drug candidates from the knowledge graph

To extract and rank drug candidates for COVID-19 drug repurposing from the generated knowledge graph, we employed three scores focusing on different characteristics and patterns in the knowledge graph: Motif scores (focuses on high-order patterns of interest); PageRank scores (focuses on connectivity between the drug node and the SARS-CoV-2 node); and Embedding scores (focuses on link existence probabilities learned from the knowledge graph). The three scores covered the mainstream techniques for measuring potential association between drug and virus nodes (i.e. local distance, global distance and learning-based distance respectively). Higher scores represent a stronger potential association between the drug and COVID-19 virus. Additionally, we explored both linear and non-linear methods to integrate the three scores and evaluated their performance (section 2.7).

2.4.1. Motif scores

Motif-based graph analysis is a classic bioinformatics technique which allows efficient extraction of target relations of interest (such as drug-virus-target linkages) from large-scale information networks (such as a COVID-19 knowledge graph). A motif, essentially a connected graph of a few nodes and edges, is often considered to be a fundamental building block of large and complex networks. Motif discovery algorithms are usually employed to identify frequent high-order patterns of interest (i.e. motifs) in knowledge graphs.[22] Motifs relevant to drug repurposing such as “drug-protein-virus” and “drug-disease/symptom-virus” were included (**Figure 3**). Subgraphs that match the motifs of interest were extracted using motif-clique discovery algorithms previously described by Hu et al.[23] Motif-clique is a dense subgraph (i.e. the connected subgraph composed by all possible motif-instances) that contains valuable information regarding an input “motif”. For example, the motif-clique in **Figure 2** shows two human protein-coding genes (NR3C1 [nuclear receptor subfamily 3, group C, member 1] corresponds to the glucocorticoid receptor which is responsible for a wide variety of effects mediating growth, metabolism, and immune response; POU1F1 [POU domain, class 1, transcription factor 1] regulates transcription of the growth hormone) are both targeted by the SARS-CoV-2, and share linkages with 34 symptoms (denoted by the symptom ID from HPO) corresponding to the motif of interest “virus-protein-symptom”. It should be noted that the motifs can also be designed by the user to customise the focus of the Motif score. Next, we adapted Jaccard coefficient to incorporate the motifs given by the user in order to compute it on the knowledge graph. The motif-based Jaccard coefficient is described as Algorithm 1, where f_{st} denotes the frequency of the motif instances that contain both *s* and *t*, and f_i denotes the frequency of the motif instances that only contain node *i*. By enumerating all the motifs in the given set *M*, the algorithm can calculate a score for the node pair (*s*, *t*). By assigning *s* as a drug and *t* as SARS-CoV-2, we can compute the motif score with respect to the set of motifs of interest.



Algorithm 1 Motif-based Jaccard Coefficient

```
Input: $G = (V, E, L), \mathcal{M}, (s, t) \in V \times V \setminus E$
Output: $P(e_{st}|\mathcal{M})$
$f_{st} \leftarrow 0,\ f_s \leftarrow 0,\ f_t \leftarrow 0$;
for $\tau \in \mathcal{M}$ do
    $M_{st} \leftarrow \{m | m \simeq \tau \& m \in G \& s \in V_m \& t \in V_m\}$;
    $M_s \leftarrow \{m | m \simeq \tau \& m \in G \& s \in V_m\}$;
    $M_t \leftarrow \{m | m \simeq \tau \& m \in G \& t \in V_m\}$;
    $f_{st} \leftarrow f_{st} + |M_{st}|,\ f_s \leftarrow f_s + |M_s|,\ f_t \leftarrow f_t + |M_t|$;
$P(e_{st}|\mathcal{M}) \leftarrow \frac{f_{st}}{f_s + f_t}$;
return $P(e_{st}|\mathcal{M})$.
```

2.4.2. PageRank scores

A drug candidate might have multiple relations interlinked with COVID-19 related genes, proteins, diseases and symptoms. This set of drug candidates were further ranked using computational scores which quantified the strength of association between each drug candidate and COVID-19 in the knowledge graph, in terms of the number and length of shared interlinkages. Due to the generality of the PageRank score (from which a variety of other scores were derived),[24] it is considered an important indicator for node ranking of the knowledge graphs and can be calculated by the function below, where *d* is the damping factor, which is usually set as 0.85.[25] For each drug *p*, *M(p)* is the set of predecessors of *p* in the knowledge graph and *L(p)* is the set of successors in the knowledge graph. *N* is the number of candidate drugs. Finally, the PageRank score of drug *p* is calculated as *PR(p)*. To apply it into large scale data, we sped up the calculation by updating *PR(p)* with interactions.

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d\sum \frac{\mathrm{PR}(p_i)}{\mathrm{L}(p_j)} \qquad (1)$$

2.4.3. Embedding scores

The COVID-19 knowledge graph is large, high dimensional, and sparse (meaning that most of the items have no linkage with one another). Knowledge graph embedding is the task of completing the knowledge graphs by probabilistically inferring the missing arcs from the existing graph structure. It projects the sparse and high dimensional graph representation vector space into a lower dimensional dense space; then the algorithm is trained to distinguish positive pairs (i.e., the node pairs with an edge in the knowledge graph) and negative pairs (i.e., the node pairs without an edge in the knowledge graph) based on the inner products of their embeddings.

Therefore, the drug repurposing problem can be reduced as the link prediction problem, which is a classic classification task in the area of machine learning. Specifically, given a drug candidate, link prediction will calculate the existential probability of the potential edge between this drug and SARS-CoV-2. Since knowledge graph embedding is the current state-of-the-art tool to fulfil this task, we defined the corresponding edge existential probability as the embedding score.

We use TransE_L2 to train the model[26] for each node in the COVID-19 knowledge graph. Specifically, given a candidate drug *x*, we predicted the existential probability (i.e. the embedding score) for *x* based on the embeddings of *x* and SARS-CoV-2, denoted as $h(x)$ and $h(y)$ respectively. Then the embedding score of drug x could be calculated by the equation as below, where *x* is the drug candidate and *y* the SARS-CoV-2 virus.

$$E\ (x) = \log(1 + \mathrm{e}^{-\mathrm{TransE\_L2}((h\ (x), h(y))}) \qquad (2)$$

Note that motif scores and PageRank scores are deterministic algorithms, i.e., the strategy is fixed regardless of the distribution of the data. The embedding score was generated from the learning algorithm which requires labelled data; in our case, the labelled data were sampled from the existing COVID-19 knowledge graph rather than the drug-virus domain knowledge. To train the embeddings, the algorithm collected part of the node pairs that are connected by an edge in the COVID-19 knowledge graph (positive samples) and node pairs without an edge (negative samples). These positive samples and negative samples did not necessarily contain drug or SARS-CoV-2 nodes and were randomly collected from the existing COVID-19 knowledge graph. Note that such labelled data (i.e., positive samples and negative samples) did not involve any drug-virus relation.

2.5. Integrated algorithm analysis

In the simplest case, the three scores can be integrated using a linear function *f(x) = α Motif(x) + β PageRank(x) + γ Embedding(x)*, where *α+β+γ=1*, *x* denotes a drug candidate for COVID-19 drug repurposing, and the three parameters (i.e., *α*, *β* and *γ*) represents the relative weighting of each score. The choices of these weights (i.e., *α*, *β* and *γ*) could be manually tuned depending on exact use case. A larger *α* would be preferred in cases where significant motifs for effective drug repurposing are well-known, or when only part of the knowledge graph are of interest (e.g. in use cases where drugs should be recommended only by their proximities with symptoms and diseases). A larger *β* could be used in cases where pathway analysis is preferred. A larger *γ* would be preferred in cases where more labelled data (i.e., the drugs that are known to be effective for COVID-19 treatment) are available, because of the powerful predicting ability of knowledge graph embedding. For the purpose of this study, we focused on reporting PageRank score (i.e. setting β=1, α=γ=0) as PageRank score does not rely on known significant motifs nor labelled data.

2.6 Evaluation

To evaluate the performance of this method on proposing drug candidates for repurposing as COVID-19 treatment, we reported the percentage of drugs proposed by our algorithm that are under or completed clinical trial for COVID-19 treatment. It should be noted that the mere fact of being under or completed clinical trial does not imply a drug's efficacy as COVID-19 treatment. Also, no true/false negative data could be inferred from clinical trials. We further calculated quantitative indicators, including Precision, Recall and F1 score, as defined below:

$$\text{Precision} = \frac{Number\ of\ drugs\ proposed\ that\ are\ under\ or\ completed\ clinical\ trial}{Total\ number\ of\ drugs\ proposed}$$

$$\text{Recall} = \frac{Number\ of\ drugs\ proposed\ that\ are\ under\ or\ completed\ clinical\ trial}{Total\ number\ of\ drugs\ under\ or\ completed\ clinical\ trial}$$

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

2.7 Exploratory analyses

We also explored automatic learning of the optimal function f(x) to integrate these three scores depicted in section 2.5, using both linear and non-linear models. For linear models, we trained a logistic regression and a linear support vector machine (LSVM) to learn the optimal parameters (i.e. α, β, γ) depicted above. For non-linear models, we trained quadratic SVM (QSVM), cubic SVM (CSVM), Gaussian SVM (GSVM); and five neural networks (NN) with different topologies, namely narrow NN (NNN, one layer with 10 neurons), middle NN (MNN, one layer with 25 neurons), wide NN (WNN, one layer with 100 neurons), duplex NN (DNN, two layers with 10x10 neurons) and triple NN (TNN, three layers with 10x10x10 neurons). We reduced the ranking problem into a binary classification problem. Specifically, we defined $F(x, y) = f(x) - f(y)$, where x and y are two drugs and f is the function to integrate the three scores. Then given two drugs x and y in the list of drugs under/completed clinical trial as described in section 2.6 with x's rank higher than y's rank, we define indicator function $I\ (F\ (x, y)>0) = 1$ & $I\ (F\ (x, y) \leqslant 0) = -1$. We then train the model by minimizing the loss function $Loss = argmin \sum_{x,y} I\ (F\ (x, y))$. We split the list of drugs under/completed clinical trial into 80% for training and 20% for validation, and conducted 5-fold cross validation to reduce potential of overfitting.

2.8. Software used

Java, MATLAB and R were used for all analysis.

3. Results

The complete knowledge graph contains over 48,000 nodes and 1,337,000 edges. The nodes are composed of 13162 diseases, 220 virus proteins/genes, 6924 viruses (strains), 10077 symptoms, 12931 host proteins/genes and 11866 drugs. We described the breakdown of the edges in Table 1. A total of 13,563 molecules in the DrugBank database were evaluated, of which 5624 molecules were identified from the knowledge graph by the drug repurposing algorithms (i.e. all three scores described above are greater than zero).

112 drug molecules had the top 2% PageRank scores, of which 50 existing oral and intravenous drugs with other FDA/EMA-approved indications were reported in the final results. The list of proposed drug candidates for COVID-19 repurposing are listed in **Table 2**. The full list of all drug candidates evaluated in the knowledge graph and scripts for drug repurposing applications are released for open access at https://github.com/Sheldon2016/covid19kg.

The proposed drug candidates include agents from a variety of drug classes, including 12 drugs for cardiovascular conditions (8 statins, moexipril, isosorbide mononitrate/dinitrate, spironolactone, bezafibrate), 11 drugs for treating infections (4 antiviral agents, 2 antiparasitic agents, 3 antibacterial agents, 2 antifungal agents), 10 hypnotics or anticonvulsants, 7 antineoplastic agents, 1 immunosuppressant, 4 hormonal agents and 4 other agents (pirfenidone, ibuprofen, amitriptyline, dexamethasone, fostamatinib). Notably, newer drugs (including remdesivir) were not ranked among the results due to lack of data for those agents in the bioinformatic data sources included in this study.

For the evaluation of the performance of our algorithm, **Figure 4** and **Table S1** shows that precision decreases and in contrary, recall and F1-score increase as we used a lower threshold for the top n% drugs to be included in our final results.

Regarding the integrated algorithm analysis, during our initial evaluation, we found that the three scores in our study are consistent in most cases (Table 2). For example, ritonavir obtained 95.8 Motif score, 100.0 PageRank score and 96.5 Embedding score. We re-scaled the three scores into the corresponding percentage of drugs that it outweighs for fair comparison which means Ritonavir outweighs 95.8% drugs in DrugBank according to $P(e_{st}|M)$ where s = ritonavir, t = SARS-CoV-2 and M is the set of motifs in Figure 3). There are also cases where the three scores are inconsistent such as eszopiclone where 99.97 PageRank score and 96.73 Embedding score, but only 26.98 Motif score were obtained. This is due to the fact that linkages of eszopiclone to SARS-CoV-2 were mainly through pathways in the protein-protein interaction network, which were not captured by the motifs in Figure 3. In our exploratory analyses, neural network models generally outperform the linear models and most SVM model, except for the more complex TNN model which had an obvious performance descent (**Figure 5**). It should be noted that the choices of the three parameters (i.e., α, β and γ) and linear or non-linear integration algorithms depend on exact use cases and parameter tuning is required. We therefore focused on reporting PageRank scores for the purpose of this study (i.e. setting β=1, α=γ=0) and in our discussion as PageRank scores do not rely on known significant motifs nor labelled data).

4. Discussion

To our knowledge, this is the most comprehensive COVID-19 knowledge graph for the purposes of drug identification for drug repurposing, with integration of major openly available bioinformatics data sources, linked with information on drug-target interactions, gene-gene interactome and gene-disease network, which has not been considered in existing computational and network-based drug repurposing studies.[27] In general, drugs shown to be useful in preliminary reports of ongoing clinical trials or hypothesised for COVID-19 treatment in previous literature were also ranked as superior in our results compared to other drugs. In addition, our results also revealed that drug candidates that were not postulated to have any effect on COVID-19 may be considered for further evaluation in clinical trials or observational studies for their effectiveness to treat COVID-19.

4.1. Anti-infective drugs

Ritonavir and lopinavir were ranked highest in our results. In a randomised trial of 199 patients with severe COVID-19, the addition of lopinavir-ritonavir (400/100 mg) twice daily for 14 days to standard care did not decrease the time to clinical improvement compared with standard care alone.[28] Yet, an open-label randomised trial showed positive results with the use of interferon beta-1b, lopinavir-ritonavir and ribavirin combination compared to lopinavir-ritonavir alone, in alleviating symptoms and shortening the duration of viral shedding and hospital stay for non-severe COVID-19.[29] Results from our knowledge graph and the current literature suggest that the role of ritonavir and lopinavir warrants further investigation, especially when used in combination with other agents.

Our results include a number of anti-infective agents besides lopinavir and ritonavir. Nelfinavir has been reported to inhibit cell fusion caused by the SARS-CoV-2 spike (S) glycoprotein and thus may possess antiviral activity against COVID-19.[30] Bictegravir had been proposed in computational analysis studies to be a 3CLpro inhibitor which may be a potential agent against SARS-CoV-2.[31] These findings were consistent with our results. In contrast, other antiviral agents currently under evaluation or in clinical trials,[32] including antivirals against influenza viruses such as oseltamivir, favipiravir and umifenovir, antivirals treating hepatitis C such as danoprevir were not proposed in our knowledge graph. Specific antibacterial agents such as azithromycin and antiparasitic agents such as ivermectin had been evaluated as treatment for COVID-19.[33] Previous studies suggest no clinical benefit for azithromycin as mono or adjunct therapy in COVID-19,[34] whereas drug levels required for ivermectin for activity against SARS-CoV-2 exceed safe drug doses in vivo.[35] In contrast, our results suggest that other antibacterial agents, specifically colistin and prufloxacin, may warrant further investigation.

4.2. Cardiovascular drugs

Statins were ranked in the top 1% of our results. Statins are known inhibitors of the MYD88 pathway, which results in marked inflammation, and have been reported to stabilise MYD88 levels in the setting of external stress in vitro and in animal studies.[36] Dysregulation of MYD88 has been noted and associated with poor outcomes in SARS-CoV and MERS-CoV infections. Statins are also known for their pleiotropic anti-inflammatory, antithrombotic and immunomodulatory effects, and have been proposed to have a potential role as adjunctive therapy to mitigate endothelial dysfunction and dysregulated inflammation in patients with COVID-19 infection.[37] However, there were reports that statins could induce ACE2 expression and thereby increase the risk of COVID-19 entry. In a retrospective observational study, involving 13,981 patients with COVID-19 in the Hubei Province China, 1,219 received statins. The 28-day all-cause mortality was 5.2% and 9.4% in the matched statin and non-statin groups, respectively, with an adjusted hazard ratio of 0.58.[38] A meta-analysis which included two retrospective studies in China, one in the United States and one in Italy showed a significantly reduced hazard for fatal or severe disease with the use of statins (pooled HR = 0.70; 95% CI 0.53-0.94).[39]

Besides statins, four other cardiovascular agents, namely moexipril, isosorbide mononitrate/dinitrate, spironolactone, bezafibrate, were also drug candidates in our results. Although Angiotensin-converting enzyme 2 (ACE2) was identified as one of the cellular receptors facilitating SARS-CoV-2 entry into host cells, ACE2 expression has also been associated with decreased severity of acute respiratory distress syndrome, which is a major complication of COVID-19 especially in severe cases; and also has a protective effect in heart failure.[40] Certain cardiovascular agents, including ACE inhibitors (ACEIs), angiotensin-II receptor blockers (ARBs), spironolactone had been shown to increase ACE2 expression in animal models.[40] Previous retrospective studies in hospitalised patients with COVID-19 in China also suggest that inpatient use of ACEI/ARB was associated with lower risk of all-cause mortality compared with ACEI/ARB nonusers.[41] Ibuprofen, a non-steroidal anti-inflammatory agent ranked at the top 2% of our results, had also shown to increase ACE2 expression and attenuate cardiac fibrosis in animal models,[42] but have not been further evaluated in other studies.

Isosorbide mononitrate or dinitrate ranked at the top 0.5% of our results. Nitric oxide had been suggested as a potential therapy in COVID-19 by countering endothelial dysfunction and nitric oxide deficiency due to COVID-19 infection and interfering with the interaction between SARS-CoV-2 and ACE-2.[43] Inhaled nitric oxide has also been under evaluation for COVID-19 in clinical trials.[44] Isosorbide mononitrate or dinitrate, an oral vasodilating agent, is converted to free radical nitric oxide endogenously and could also be potentially beneficial in patients with COVID-19. Bezafibrate ranked second last in our proposed drug candidates. Fibrates have demonstrated anticoagulant and cardiovascular protective effects in patients with metabolic syndrome,[45] with potential protective effects on kidney function,[46] which may offer benefit in patients with complications due to COVID-19 infection. While fenofibrate is being evaluated in clinical trial,[47] bezafibrate may also warrant further investigation.

4.3. Other drugs

Pirfenidone was ranked in the top 0.5% of our results. Pirfenidone is indicated for treatment of idiopathic pulmonary fibrosis and proposed to be beneficial for acute lung injury and acute respiratory distress syndrome in severe cases of COVID-19.[48] Clinical trials are underway to evaluate its efficacy in these cases.[49]

Hormones and hormonal agents, including estradiol, progesterone, tamoxifen (a selective estrogen receptor modulator) and mifepristone (an anti-progestogen) were ranked in the top 2% of our results. Endogenous hormones, estradiol and progesterone, exert a wide array of effects in both men and women. In the context of COVID-19, their immunomodulatory and anti-inflammatory effects have been of interest.[50] High physiological concentrations of 17β-estradiol and progesterone favour a state of decreased innate immune inflammatory response while enhancing immune tolerance and antibody production, which in turn is suggested to potentially improve immune dysregulation and prevent cytokine storm caused by COVID-19 infections.[50] Indeed, exogenous estrogen and progesterone therapy and tamoxifen have been under evaluation in clinical trials.[51]

Dexamethasone was also ranked in the top 2% of all drugs, and hydrocortisone and prednisone were also ranked in the top 5%. Data from randomised trials overall support the role of glucocorticoids for severe COVID-19. From a meta-analysis of seven trials which included 1803 critically ill patients with COVID-19, glucocorticoids reduced 28-day mortality compared with standard care or placebo (32 versus 40 percent, odds ratios [OR] 0.66, 95% CI 0.53-0.82) and were not associated with an increased risk of severe adverse events.[6] As a result, dexamethasone was recommended by the WHO for severely ill patients with COVID-19 who are on supplemental oxygen or ventilatory support, replaceable by other glucocorticoids at equivalent doses,[6] but they were not recommended for prevention in non-severe cases because of potential adverse effects.

Our results also included a number of anticonvulsants, hypnotics, antineoplastic agents including tyrosine kinase inhibitors and cytotoxic agents such as taxanes. These agents may have been proposed by the knowledge graph algorithms due to their broad effect on a larger number of linkages with various endogenous signalling pathways, genes and proteins which shared common linkages with COVID-19 and other viruses. However, these agents are unlikely to be proposed for the treatment of COVID-19 due to their severe adverse effect profile, including cytotoxicity, immunosuppression and respiratory depression, that could potentially worsen patient outcomes over any potential beneficial effect against COVID-19.

Our study has limitations. Notably, study results serve to generate hypotheses on which existing drugs may have greater potential to be repurposed for COVID-19 treatment. Yet it does not provide any clinical or biological evidence on the effectiveness or mechanisms of action for the proposed drug candidates in treating COVID-19, which needs further validation and evaluation in future clinical trials or observational studies. Further, while our results proposed potential drug candidates for drug repurposing, this information must be interpreted alongside the drugs' adverse effect profile and practicality for use in patients with COVID-19 to ensure that any potential benefit outweighs known adverse effects. The toxicity of drugs were not evaluated using the knowledge graph in this study. We refer interested readers to the current biomedical literature for a detailed review of the toxicity alongside adverse effect profiles of the proposed drug candidates. Some new drugs were not included due to the lack of data in the data sources at the time of this study. Combinations of drugs that may be candidates for COVID-19 repurposing remains to be explored. Currently, tuning parameters in the integrated scoring algorithm described above require code modification which may not yet be user-friendly to healthcare professionals or other researchers. In future work, we could design a publicly accessible user interface as well as an automatic parameter tuning method to assist the user with selecting optimal parameters.

5. Conclusions

We developed a COVID-19 knowledge graph from large-scale bioinformatic databases for drug repurposing. Using an integrated algorithm to integrate three computation scores, a set of 50 drug candidates were shortlisted as potential treatments for COVID-19. These candidates included drugs for cardiovascular diseases, anti-infective agents, hormonal agents and steroids, among other drug classes. Some of these candidates have also been undergoing evaluation in clinical trials, while others have received relatively little attention to date. Our findings serve to generate hypotheses and prioritise drug candidates for further evaluation in clinical trials and observational studies.

Funding

This study was supported through funding from the Li Ka Shing Faculty of Medicine, The University of Hong Kong.

Contributors

EWC, RCKC, RBL, VKCY, XDL and MO conceptualized the study. EWC, RCKC and RBL provided resources and supervised the study. VKCY contributed domain knowledge on pharmacology and conducted literature review and validation. XDL and XXY curated the data. XDL and MO provided the software and conducted the formal analysis. VKCY and XXY wrote the original draft. All authors critically reviewed and commented on all other drafts.

Acknowledgements

We thank Prof. Chuan Wu and Prof. Kevin Chang for their valuable input on the concept and design of this study. We also thank Ms Lisa Y Lam for proofreading the manuscript. Vincent KC Yan, Xiaodong Li and Xuxiao Ye contributed equally to this work. Esther W Chan and Reynold CK Cheng share senior authorship.

Conflict of interests

EWC has received honorarium from the Hospital Authority and research funding from The Hong Kong Research Grants Council, The Research Fund Secretariat of the Food and Health Bureau, Narcotics Division of the Security Bureau of HKSAR, Hong Kong; National Natural Science Fund of China, China; Wellcome Trust, United Kingdom; Bayer, Bristol-Myers Squibb, Pfizer, and Takeda, for work unrelated to this study. ICKW has received research funding outside the submitted work from the Hong Kong Research Grants Council and the

Hong Kong Health and Medical Research Fund, National Institute for Health Research in the UK, European Commission, Amgen, Bayer, Bristol-Myers Squibb, GSK, and Janssen. All other authors declare no competing interests.

Data sharing

The COVID-19 knowledge graph and scripts for drug repurposing applications generated from this work will be released with publication for open access at https://github.com/Sheldon2016/covid19kg.

References

[1] World Health Organisation. WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/. accessed **2020**.

[2] US FDA. Emergency Use Authorization (EUA) of the Pfizer-BioNTech COVID-19 Vaccine to Prevent Coronavirus. Fact sheet for healthcare providers administering vaccine. https://www.fda.gov/media/144413/download. accessed **2021**.

[3] US FDA. Emergency Use Authorization (EUA) of the Moderna COVID-19 Vaccine to prevent Coronavirus Disease 2019 (COVID-19). Factsheet for healthcare providers administering vaccine. https://www.fda.gov/media/144637/download?utm_medium=email&utm_source=govdelivery. accessed **2020**.

[4] US FDA. Emergency use authorization (EUA) of the Janssen COVID-19 vaccine to prevent coronavirus disease 2019 (COVID-19). https://www.fda.gov/media/146304/download. accessed **2021**.

[5] US CDC. Joint CDC and FDA Statement on Johnson & Johnson COVID-19 Vaccine. https://www.cdc.gov/media/releases/2021/s0413-JJ-vaccine.html. accessed **2021**.

[6] F. Lamontagne, T. Agoritsas, H. Macdonald, Y. S. Leo, J. Diaz, A. Agarwal, J. A. Appiah, Y. Arabi, L. Blumberg, C. S. Calfee, B. Cao, M. Cecconi, G. Cooke, J. Dunning, H. Geduld, P. Gee, H. Manai, D. S. Hui, S. Kanda, L. Kawano-Dourado, Y. J. Kim, N. Kissoon, A. Kwizera, J. H. Laake, F. R. Machado, N. Qadir, R. Sarin, Y. Shen, L. Zeng, R. Brignardello-Petersen, L. Lytvyn, R. Siemieniuk, D. Zeraatkar, J. Bartoszko, L. Ge, B. Maguire, B. Rochwerg, G. Guyatt, P. O. Vandvik, *BMJ* **2020**, *370*, m3379.

[7] M. Mahévas, V.-T. Tran, M. Roumier, A. Chabrol, R. Paule, C. Guillaud, E. Fois, R. Lepeule, T.-A. Szwebel, F.-X. Lescure, F. Schlemmer, M. Matignon, M. Khellaf, E. Crickx, Terrier, *BMJ* **2020**, *369*.

[8] B. Li, R. Cheng, J. Hu, Y. Fang, M. Ou, R. Luo, K. Chang, X. Lin, *MC-Explorer: analyzing and visualizing motif-cliques on large networks,* 2020.

[9] F. Cheng, *Methods Mol. Biol.* **2019**, *1878*, 243.

[10] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Hüttenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. McGregor, Q. Li, B. Meyer, F. Roesch, T. Vallet, A. Mac Kain, L. Miorin, E. Moreno, Z. Z. C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lyu, C. J. P. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Rakesh, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, X.-P. Huang, Y. Liu, S. A. Wankowicz, M. Bohn, M. Safari, F. S. Ugur, C. Koh, N. S. Savar, Q. D. Tran, D. Shengjuler, S. J. Fletcher, M. C. O'Neal, Y. Cai, J. C. J. Chang, D. J. Broadhurst, S. Klippsten, P. P. Sharp, N. A. Wenzell, D. Kuzuoglu-Ozturk, H.-Y. Wang, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R. M. Stroud, A. D. Frankel, O. S. Rosenberg, K. A. Verba, D. A. Agard, M. Ott, M. Emerman, N. Jura, M. von Zastrow, E. Verdin, A. Ashworth, O. Schwartz, C. d'Enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. N. Floor, J. S. Fraser, J. D. Gross, A. Sali, B. L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. García-Sastre, K. M. Shokat, B. K. Shoichet, N. J. Krogan, *Nature* **2020**, *583*, 459.

[11] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, *Cell Discovery* **2020**, *6*, 14.

[12] S. Dotolo, A. Marabotti, A. Facchiano, R. Tagliaferri, *Brief. Bioinform.* **2020**.

[13] B. Li, R. Cheng, J. Hu, Y. Fang, M. Ou, R. Luo, K. C. C. Chang, X. Lin., in *IEEE 36th International Conference on Data Engineering,* 2020, 1722.

[14] Y. Jin, W. Zhang, X. He, X. Wang, X. Wang, in *IEEE 36th International Conference on Data Engineering*, 2020, 145.

[15] D. Domingo-Fernández, S. Baksi, B. Schultz, Y. Gadiya, R. Karki, T. Raschka, C. Ebeling, M. Hofmann-Apitius, A. T. Kodamullil, *Bioinformatics* **2020**.

[16] The CovidGraph Project. CovidGraph. https://covidgraph.org/. accessed **2020**.

[17] D. M. Gysi, Í. D. Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, S. D. Ghiassian, J. J. Patten, R. Davey, J. Loscalzo, A.-L. Barabási, *arXiv* **2020**, *2004.07229*.

[18] Z. Wang, J. Zhang, J. Feng, Z. Chen, in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI Press, Québec City, Québec, Canada 2014, 1112.

[19] UniProt Consortium, *Nucleic Acids Res.* **2014**, *42*, D191.

[20] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, *Nucleic Acids Res.* **2019**, *47*, D955.

[21] National Library of Medicine. Medical Subject Headings. https://www.nlm.nih.gov/mesh/meshhome.html. accessed **2020**.

[22] C. Ma, R. Cheng, L. V. S. Lakshmanan, T. Grubenmann, Y. Fang, X. Li, *Proc. VLDB Endow.* **2019**, *13*, 155.

[23] J. Hu, R. Cheng, K. C. Chang, A. Sankar, Y. Fang, B. Y. H. Lam, "Discovering maximal motif cliques in large heterogeneous information networks", presented at *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019.

[24] X. Li, R. Cheng, K. Chang, C. Shan, C. Ma, H. Cao, *Proc. VLDB Endow.* **2021**, *14*.

[25] X. Li, T. N. Chan, R. Cheng, C. Shan, C. Ma, K. Chang, "Motif paths: a new approach for analyzing higher-order semantics between graph nodes", 2019.

[26] D. Zheng, X. Song, C. Ma, Z. Tan, Z. Ye, J. Dong, H. Xiong, Z. Zhang, G. Karypis, *arXiv* **2020**, *2004.08532*.

[27] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, *Cell discovery* **2020**, *6*, 14.

[28] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai, M. Wei, X. Li, J. Xia, N. Chen, J. Xiang, T. Yu, T. Bai, X. Xie, L. Zhang, C. Li, Y. Yuan, H. Chen, H. Li, H. Huang, S. Tu, F. Gong, Y. Liu, Y. Wei, C. Dong, F. Zhou, X. Gu, J. Xu, Z. Liu, Y. Zhang, H. Li, L. Shang, K. Wang, K. Li, X. Zhou, X. Dong, Z. Qu, S. Lu, X. Hu, S. Ruan, S. Luo, J. Wu, L. Peng, F. Cheng, L. Pan, J. Zou, C. Jia, J. Wang, X. Liu, S. Wang, X. Wu, Q. Ge, J. He, H. Zhan, F. Qiu, L. Guo, C. Huang, T. Jaki, F. G. Hayden, P. W. Horby, D. Zhang, C. Wang, *N. Engl. J. Med.* **2020**, *382*, 1787.

[29] I. F. Hung, K. C. Lung, E. Y. Tso, R. Liu, T. W. Chung, M. Y. Chu, Y. Y. Ng, J. Lo, J. Chan, A. R. Tam, H. P. Shum, V. Chan, A. K. Wu, K. M. Sin, W. S. Leung, W. L. Law, D. C. Lung, S. Sin, P. Yeung, C. C. Yip, R. R. Zhang, A. Y. Fung, E. Y. Yan, K. H. Leung, J. D. Ip, A. W. Chu, W. M. Chan, A. C. Ng, R. Lee, K. Fung, A. Yeung, T. C. Wu, J. W. Chan, W. W. Yan, W. M. Chan, J. F. Chan, A. K. Lie, O. T. Tsang, V. C. Cheng, T. L. Que, C. S. Lau, K. H. Chan, K. K. To, K. Y. Yuen, *Lancet* **2020**, *395*, 1695.

[30] F. Musarrat, V. Chouljenko, A. Dahal, R. Nabi, T. Chouljenko, S. D. Jois, K. G. Kousoulas, *J. Med. Virol.* **2020**.

[31] W. Guan, W. Lan, J. Zhang, S. Zhao, J. Ou, X. Wu, Y. Yan, J. Wu, Q. Zhang, *Virol. Sin.* **2020**.

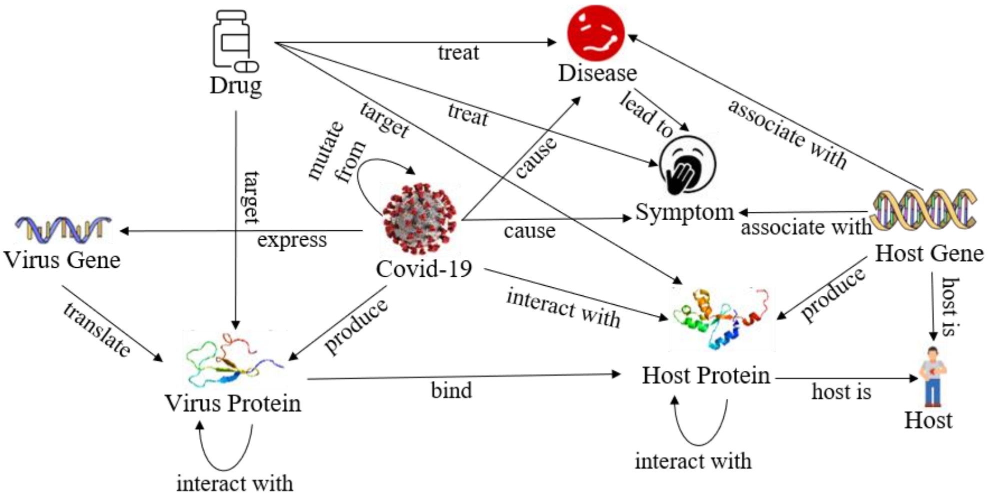
[32] S. Lam, A. Lombardi, A. Ouanounou, *Eur. J. Pharmacol.* **2020**, *886*, 173451.

[33] G. Magro, *Virus Res.* **2020**, *286*, 198070.

[34] A. B. Cavalcanti, F. G. Zampieri, R. G. Rosa, L. C. P. Azevedo, V. C. Veiga, A. Avezum, L. P. Damiani, A. Marcadenti, L. Kawano-Dourado, T. Lisboa, D. L. M. Junqueira, E. S. P. G. M. de Barros, L. Tramujas, E. O. Abreu-Silva, L. N. Laranjeira, A. T. Soares, L. S. Echenique, A. J. Pereira, F. G. R. Freitas, O. C. E. Gebara, V. C. S. Dantas, R. H. M. Furtado, E. P. Milan, N. A. Golin, F. F. Cardoso, I. S. Maia, C. R. Hoffmann Filho, A. P. M. Kormann, R. B. Amazonas, M. F. Bocchi de Oliveira, A. Serpa-Neto, M. Falavigna, R. D. Lopes, F. R. Machado, O. Berwanger, *N. Engl. J. Med.* **2020**, *383*, 2041.

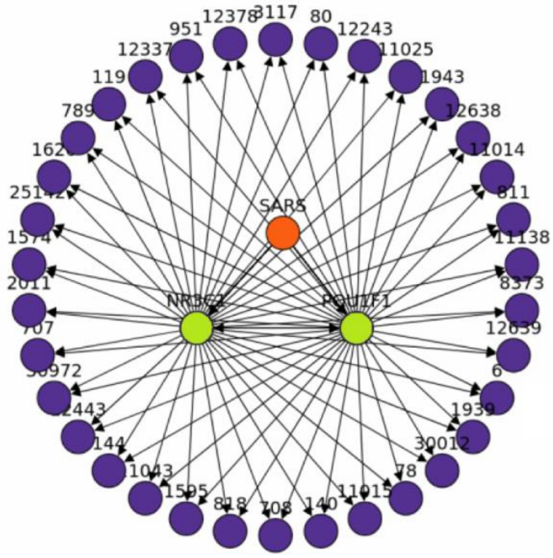
[35] F. Heidary, R. Gharebaghi, *J. Antibiot. (Tokyo)* **2020**, *73*, 593.
[36] S. Yuan, *mBio* **2015**, *6*, e01120.
[37] K. C. H. Lee, D. W. Sewa, G. C. Phua, *Int. J. Infect. Dis.* **2020**, *96*, 615.
[38] X. J. Zhang, J. J. Qin, X. Cheng, L. Shen, Y. C. Zhao, Y. Yuan, F. Lei, M. M. Chen, H. Yang, L. Bai, X. Song, L. Lin, M. Xia, F. Zhou, J. Zhou, Z. G. She, L. Zhu, X. Ma, Q. Xu, P. Ye, G. Chen, L. Liu, W. Mao, Y. Yan, B. Xiao, Z. Lu, G. Peng, M. Liu, J. Yang, L. Yang, C. Zhang, H. Lu, X. Xia, D. Wang, X. Liao, X. Wei, B. H. Zhang, X. Zhang, J. Yang, G. N. Zhao, P. Zhang, P. P. Liu, R. Loomba, Y. X. Ji, J. Xia, Y. Wang, J. Cai, J. Guo, H. Li, *Cell Metab.* **2020**, *32*, 176.
[39] C. S. Kow, S. S. Hasan, *Am. J. Cardiol.* **2020**, *134*, 153.
[40] M. Gheblawi, K. Wang, A. Viveiros, Q. Nguyen, J.-C. Zhong, A. J. Turner, M. K. Raizada, M. B. Grant, G. Y. Oudit, *Circ. Res.* **2020**, *126*, 1456.
[41] P. Zhang, L. Zhu, J. Cai, F. Lei, J. J. Qin, J. Xie, Y. M. Liu, Y. C. Zhao, X. Huang, L. Lin, M. Xia, M. M. Chen, X. Cheng, X. Zhang, D. Guo, Y. Peng, Y. X. Ji, J. Chen, Z. G. She, Y. Wang, Q. Xu, R. Tan, H. Wang, J. Lin, P. Luo, S. Fu, H. Cai, P. Ye, B. Xiao, W. Mao, L. Liu, Y. Yan, M. Liu, M. Chen, X. J. Zhang, X. Wang, R. M. Touyz, J. Xia, B. H. Zhang, X. Huang, Y. Yuan, R. Loomba, P. P. Liu, H. Li, *Circ. Res.* **2020**, *126*, 1671.
[42] W. Qiao, C. Wang, B. Chen, F. Zhang, Y. Liu, Q. Lu, H. Guo, C. Yan, H. Sun, G. Hu, X. Yin, *Cardiology* **2015**, *131*, 97.
[43] S. J. Green, *Microbes and infection* **2020**, *22*, 149.
[44] ClinicalTrials.gov [Internet], Bethesda (MD): National Library of Medicine (US). Search results: Trials in patients with COVID-19 and nitric oxide as intervention. https://clinicaltrials.gov/ct2/results?cond=Covid19&intr=nitric+oxide. accessed **2020**.
[45] D. Wang, B. Liu, W. Tao, Z. Hao, M. Liu, *Cochrane Database Syst. Rev.* **2015**, *2015*, Cd009580.
[46] A. Kilicarslan, B. Yavuz, G. S. Guven, E. Atalar, L. Sahiner, Y. Beyazit, M. Kekilli, N. Ozer, G. Oz, I. C. Haznedaroglu, T. Sozen, *Blood Coagul. Fibrinolysis* **2008**, *19*, 310.
[47] ClinicalTrials.gov [Internet], Bethesda (MD): National Library of Medicine (US). Identifier NCT04517396, FEnofibRate as a Metabolic INtervention for COVID-19. https://clinicaltrials.gov/ct2/show/NCT04653831. accessed **2020**.
[48] P. M. George, A. U. Wells, R. G. Jenkins, *The Lancet Respiratory Medicine* **2020**, *8*, 807.
[49] ClinicalTrials.gov [Internet], Bethesda (MD): National Library of Medicine (US). Identifier NCT04653831, Treatment with pirfenidone for COVID-19 related severe ARDS. https://clinicaltrials.gov/ct2/show/NCT04653831. accessed **2020**.
[50] F. Mauvais-Jarvis, S. L. Klein, E. R. Levin, *Endocrinology* **2020**, *161*, bqaa127.
[51] ClinicalTrials.gov [Internet], Bethesda (MD): National Library of Medicine (US). Search results: Trials in patients with COVID-19 and estrogen or progesterone as intervention. https://clinicaltrials.gov/ct2/results?cond=Covid19&intr=estrogen+OR+progesterone. accessed **2020**.

Figure 1. Structure of the COVID-19 knowledge graph



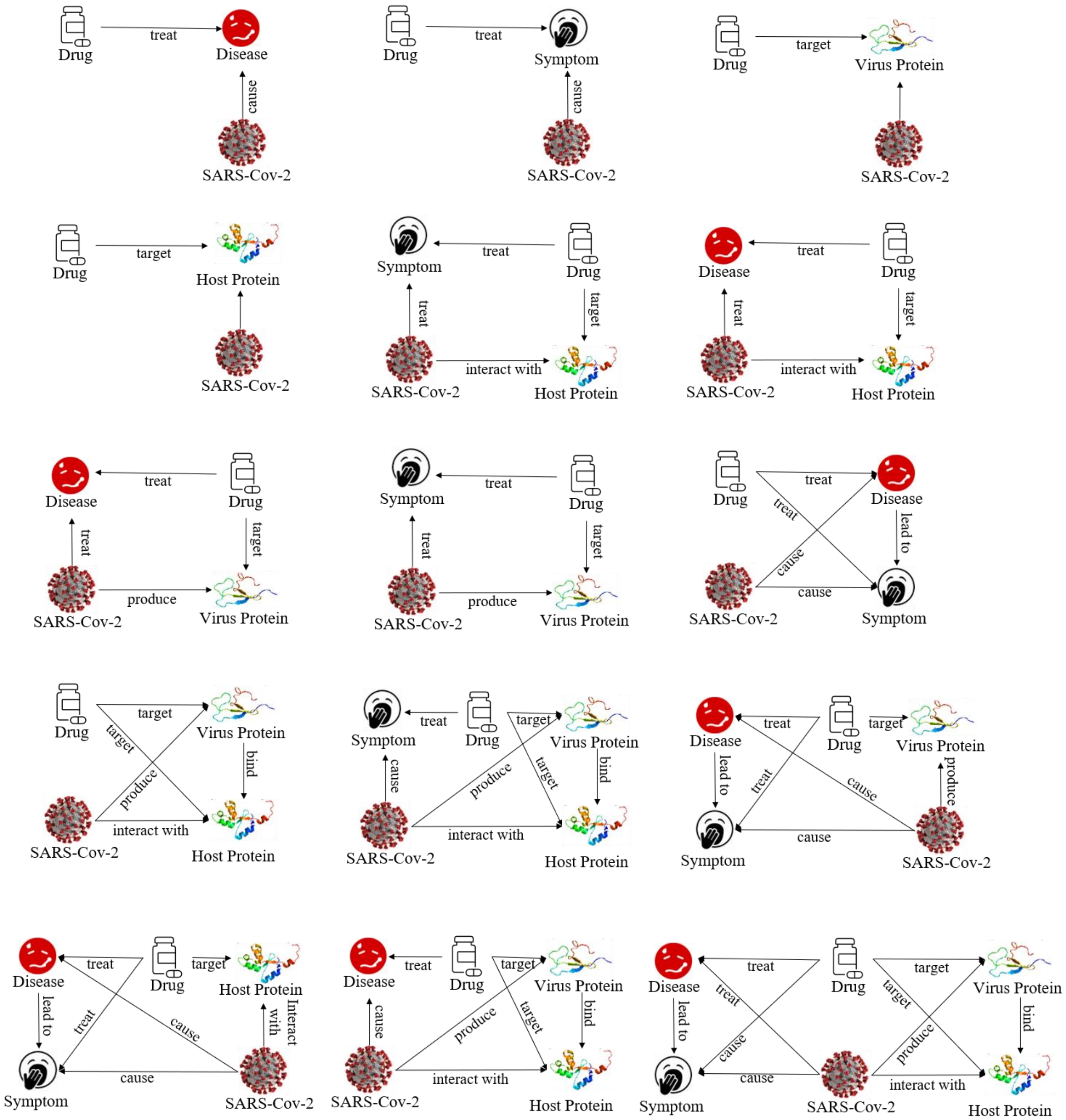
Visual schematic of the COVID-19 knowledge graph in this study. A knowledge graph is a multi-relational graph composed of entities (nodes) and relations (edges). Each node represents a specific protein, gene, drug, virus, disease or symptom, whereas each edge represents a known existing linkage between any two nodes. Data on linkages from different data sources were processed into the corresponding nodes and edges.

Figure 2. Example of motif-clique “virus-protein-symptom”



The motif-clique shown consists of 2 human proteins (green circles: NR3C1 and POU1F1) both targeted by a virus (orange circle: SARS-CoV-2) and share linkages with 34 symptoms (purple circles: annotated by symptom ID from HPO). This is one of the motif-cliques extracted from the knowledge graph using motif-discovery algorithms and corresponds to a motif of interest prespecified by the user (in this case, the “virus-protein-symptom” motif)

Figure 3. Motifs-of-interest for drug repurposing used in this study



A motif, essentially a connected graph of a few nodes and edges, is a fundamental building block of large and complex knowledge graphs. Motifs-of-interest are defined depending on the use case (e.g. drug repurposing in our study). After defining the relevant motifs-of-interest, motif-clique discovery algorithms are used to extract subgraphs that match the motifs of interest. Note each type of node only appears once in each motif for better efficiency.

Figure 4. Performance of the knowledge graph drug repurposing algorithm used in this study.

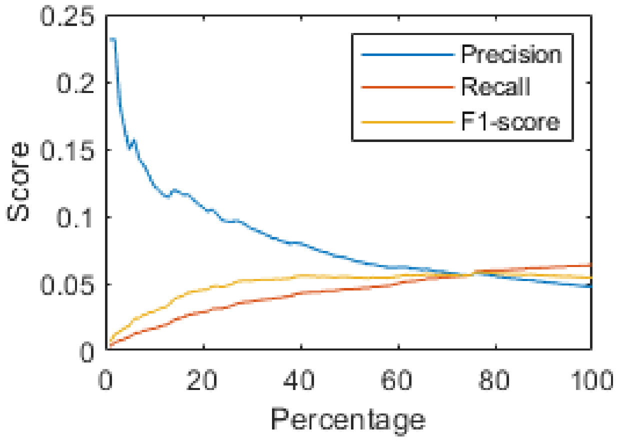


Figure 5. Accuracy of linear models (LR and LSVM) and non-linear models (SVMs except LSVM, and all NNs) used for integrating motif, PageRank and embedding scores. Models are order by increasing complexity from left to right.

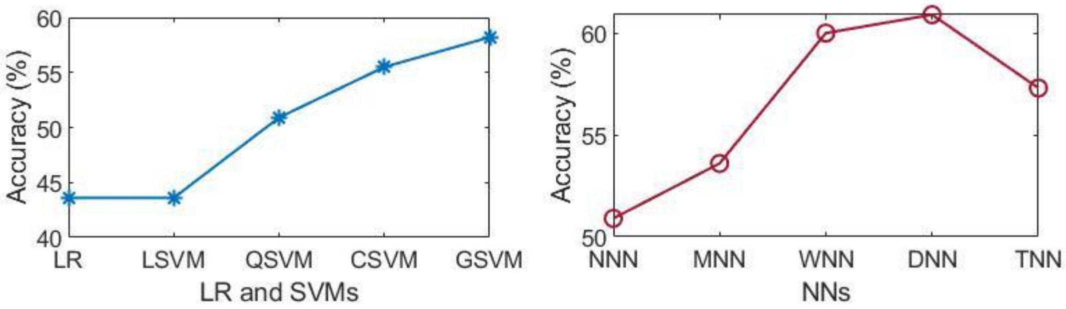


Table 1. Data sources used for inferring edges in the COVID-19 knowledge graph

Edges	**Data sources**	**Size**
Drug – Virus Protein	OpenKG	20
Drug – Disease	HPO, DrugBank	2335
Drug – Symptom	HPO, DrugBank	11730
Drug - Host Protein	DrugBank, NCBI	13749
Disease – Symptom	HPO	187342
Host Gene – Host Protein	NCBI, Literature[27]	12931
Host Gene – Disease	Disgenet	93044
Host Gene – Symptom	HPO	830344
Host Protein – Host Protein	Uniprot, Biogrid	169222
Virus Protein – Virus Protein	Biogrid	47
Virus Protein – Host Protein	OpenKG	8292
Virus - Virus	NCBI	6791
Virus – Disease	OpenKG, HPO	23
Virus – Symptom	OpenKG, HPO	70
Virus – Host Protein	Literature[27]	130
Virus – Virus Protein	OpenKG	525
Virus – Virus Gene	OpenKG	525
Virus Gene – Virus Protein	OpenKG	525

Size refers to the number of edges (representing a specific type of linkage) in the knowledge graph that were inferred from the corresponding data sources. Details of the data sources were described in Supporting Information.

Table 2. List of drug candidates for COVID-19 repurposing proposed by knowledge graph

Drugs	**Drug class**	**Motif score**	**PageRank score**	**Embedding score**
Ritonavir	Antiretroviral agent, protease inhibitor	95.80	100.00	96.51
Lopinavir	Antiretroviral agent, protease inhibitor	95.79	99.99	96.71
Pitavastatin	Lipid-modifying agent, statin	95.64	99.98	92.64
Eszopiclone	Hypnotic	26.98	99.97	96.73
Zopiclone	Hypnotic	89.98	99.97	91.84
Perampanel	Anticonvulsant, AMPA glutamate receptor antagonist	30.59	99.96	90.66
Praziquantel	Anthelmintic agent	91.16	99.95	96.64
Colistin	Antibiotic	93.29	99.94	99.44
Bictegravir	Antiviral agent, integrase inhibitor	15.43	99.93	95.56
Nelfinavir	Antiretroviral agent, protease inhibitor	89.46	99.92	93.36
Prulifloxacin	Antibiotic, fluoroquinolone	14.65	99.92	96.76
Cyclosporine	Immunosuppressant, calcineurin inhibitor	8.15	99.91	99.85
Fostamatinib	Spleen tyrosine kinase inhibitor	97.24	99.90	81.93
Moexipril	Antihypertensive agent, angiotensin-converting enzyme inhibitor	94.24	99.89	90.33

Pirfenidone	Antifibrotic agent	59.72	99.85	89.50
Isosorbide	Antianginal agent, vasodilator	26.44	99.81	52.64
Bosutinib	Antineoplastic agent, tyrosine kinase inhibitor	49.20	99.80	48.74
Dasatinib	Antineoplastic agent, tyrosine kinase inhibitor	96.60	99.73	97.25
Docetaxel	Antineoplastic agent, taxane	89.56	99.68	97.55
Lovastatin	Lipid-modifying agent, statin	95.73	99.65	96.45
Simvastatin	Lipid-modifying agent, statin	95.71	99.65	98.72
Atorvastatin	Lipid-modifying agent, statin	95.74	99.64	91.08
Flucytosine	Antifungal agent	95.69	99.60	63.87
Cerivastatin	Lipid-modifying agent, statin	95.70	99.58	93.28
Fluvastatin	Lipid-modifying agent, statin	95.69	99.57	93.80
Oxamniquine	Anthelmintic agent	95.65	99.55	81.91
Pravastatin	Lipid-modifying agent, statin	95.68	99.54	96.54
Rosuvastatin	Lipid-modifying agent, statin	95.72	99.54	94.77
Miconazole	Antifungal agent, imidazole	90.72	99.49	96.37
Ibuprofen	Nonsteroidal anti-inflammatory drug	98.40	99.48	80.73
Ponatinib	Antineoplastic agent, tyrosine kinase inhibitor	30.44	99.47	90.64
Estradiol	Hormonal agent, estrogen	93.46	99.41	99.68
Cannabidiol	Anticonvulsant, cannabinoid	29.12	99.39	85.54
Pentobarbital	Anticonvulsant, barbiturate	51.68	99.37	43.95

Amitriptyline	Antidepressant, tricyclic antidepressant	99.44	99.36	97.29
Progesterone	Hormonal agent, progestin	97.29	99.34	99.34
Temazepam	Hypnotic, benzodiazepine	88.50	99.27	92.92
Triazolam	Hypnotic, benzodiazepine	92.50	99.26	96.92
Zonisamide	Anticonvulsant	92.40	99.24	28.34
Regorafenib	Antineoplastic agent, tyrosine kinase inhibitor	30.48	99.22	93.37
Spironolactone	Antihypertensive, aldosterone receptor antagonist	97.19	99.20	98.92
Rifampicin	Antibiotic	91.26	99.18	98.60
Dexamethasone	Anti-inflammatory agent, corticosteroid	97.14	99.17	99.97
Tamoxifen	Hormonal agent, selective estrogen receptor modulator	94.37	99.13	98.96
Mifepristone	Hormonal agent, antiprogestin	97.23	99.12	95.30
Clonazepam	Anticonvulsant, benzodiazepine	91.08	99.11	99.39
Eribulin	Antineoplastic agent, microtubule inhibitor	30.69	99.07	88.32
Paclitaxel	Antineoplastic agent, taxane	52.66	99.02	85.58
Diazepam	Anticonvulsant, benzodiazepine	40.36	98.29	25.30
Bezafibrate	Lipid-modifying agent, fibrate	34.65	98.06	81.88

The proposed list of drug candidates comprises 50 existing oral and intravenous drugs with other FDA/EMA-approved indications that had top 2% PageRank scores among all ranked molecules.

Table of Contents

Knowledge graph can be used to identify potential drug candidates for repurposing as treatment of COVID-19 by using motif discovery algorithms to identify frequent high-order patterns of interest. Our findings serve to generate hypotheses and prioritise drug candidates for further evaluation in clinical trials and be further adapted for drug repurposing in other disease areas.