

OPEN ACCESS

EDITED BY

Reza Lashgari,
Shahid Beheshti University, Iran

REVIEWED BY

Dezso Modos,
Quadram Institute, United Kingdom
Yash Gupta,
Mayo Clinic Florida, United States

*CORRESPONDENCE

Tanvir Alam
✉ talam@hbku.edu.qa

SPECIALTY SECTION

This article was submitted to
Infectious Diseases: Epidemiology and
Prevention,
a section of the journal
Frontiers in Public Health

RECEIVED 16 December 2022

ACCEPTED 07 February 2023

PUBLISHED 06 March 2023

CITATION

Basit SA, Qureshi R, Musleh S, Guler R,
Rahman MS, Biswas KH and Alam T (2023)
COVID-19Base v3: Update of the
knowledgebase for drugs and biomedical
entities linked to COVID-19.
Front. Public Health 11:1125917.
doi: 10.3389/fpubh.2023.1125917

COPYRIGHT

© 2023 Basit, Qureshi, Musleh, Guler, Rahman,
Biswas and Alam. This is an open-access article
distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

COVID-19Base v3: Update of the knowledgebase for drugs and biomedical entities linked to COVID-19

Syed Abdullah Basit¹, Rizwan Qureshi¹, Saleh Musleh¹,
Reto Guler^{2,3,4}, M. Sohel Rahman⁵, Kabir H. Biswas⁶ and
Tanvir Alam^{1*}

¹College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar, ²International Centre for Genetic Engineering and Biotechnology (ICGEB), Cape Town Component, University of Cape Town, Cape Town, South Africa, ³Department of Pathology, Division of Immunology and South African Medical Research Council (SAMRC) Immunology of Infectious Diseases, Institute of Infectious Diseases and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, ⁴Wellcome Centre for Infectious Diseases Research in Africa, Institute of Infectious Diseases and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, ⁵Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, ⁶College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar

COVID-19 has taken a huge toll on our lives over the last 3 years. Global initiatives put forward by all stakeholders are still in place to combat this pandemic and help us learn lessons for future ones. While the vaccine rollout was not able to curb the spread of the disease for all strains, the research community is still trying to develop effective therapeutics for COVID-19. Although Paxlovid and remdesivir have been approved by the FDA against COVID-19, they are not free of side effects. Therefore, the search for a therapeutic solution with high efficacy continues in the research community. To support this effort, in this latest version (v3) of COVID-19Base, we have summarized the biomedical entities linked to COVID-19 that have been highlighted in the scientific literature after the vaccine rollout. Eight different topic-specific dictionaries, i.e., gene, miRNA, lncRNA, PDB entries, disease, alternative medicines registered under clinical trials, drugs, and the side effects of drugs, were used to build this knowledgebase. We have introduced a BLSTM-based deep-learning model to predict the drug-disease associations that outperforms the existing model for the same purpose proposed in the earlier version of COVID-19Base. For the very first time, we have incorporated disease-gene, disease-miRNA, disease-lncRNA, and drug-PDB associations covering the largest number of biomedical entities related to COVID-19. We have provided examples of and insights into different biomedical entities covered in COVID-19Base to support the research community by incorporating all of these entities under a single platform to provide evidence-based support from the literature. COVID-19Base v3 can be accessed from: <https://covidbase-v3.vercel.app/>. The GitHub repository for the source code and data dictionaries is available to the community from: <https://github.com/91Abdullah/covidbasev3.0>.

KEYWORDS

SARS-CoV-2, COVID-19, deep learning, machine learning, COVID-19

1. Introduction

The global COVID-19 pandemic led to millions of deaths and imposed a huge burden on the economy worldwide (1). The vaccine rollout was not able to curb the fast spread of all COVID-19 variants (2), and vaccine availability is still far from what is required (3). As a result, there are still a considerable number of people infected with this novel coronavirus SARS-CoV-2 worldwide, which suggests that equal access to vaccines across the world needs to be reinforced, particularly in low- to middle-income countries (4). Therefore, there is an urgent need to repurpose well-tolerated existing drugs with high efficacy that can be easily administered to curb the disease progression of COVID-19. Hundreds of drugs have already been tested on COVID-19 patients, mostly in hospitals, resulting in a large amount of data of varying quality (5). The guidance issued globally followed this data to some extent; for example, the FDA recommended dexamethasone for hospitalized patients who require oxygen or mechanical ventilation, and the use of tocilizumab further improves survival. Recently, the FDA has approved the use of Paxlovid™ (nirmatrelvir and ritonavir) as a prescribed drug for COVID-19; however, it is not free from side effects and concomitant use of Paxlovid™ with statins may result in an adverse drug interaction (6). Therefore, the scientific community is still looking for effective therapeutic treatment options for COVID-19.

There are only a few studies available regarding the automated searching of COVID-19 drugs. The Chinese Antibody Society launched the “COVID-19 Antibody Therapeutics Tracker” (also called “Tracker”) program in March 2020 to track antibody-based COVID-19 intervention plans in preclinical and clinical development (7). This tracker presents the analysis and visualization of COVID-19 antibody development for over 150 research programs as of the 8th August 2020. The collected data is categorized by different targets and development status, as well as country of origin. Various public domain resources, such as published literature, preprints, official websites, news feeds, social media, and government databases, were also used for data collection.

A user-friendly platform (CO-19 PDB; COVID-19 Pandemic Data Base) for COVID-19 research was developed by Ullah et al. (8). A total of 59 databases were gathered into the CO-19 PDB database between December 2019 and May 2021 and the data was organized into six different categories, namely databases for (a) digital images, (b) genomic information, (c) literature collection, (d) visualization tools, (e) chemical structure, and (f) social science-related information. These databases concentrate on extracting a variety of information, such as genomic sequences, images, the latest news updates, reports, articles, and books. Yang et al. developed the “COVID-19 Antibody Therapeutics Tracker” to track antibody-based preclinical and clinical interventions globally to combat COVID-19 (7). The authors mentioned that two antibodies, CD6 and IL-6R, have been approved for the drugs itolizumab and levilimab, respectively, and 217 antibodies are in different phases of clinical trials. The authors stopped tracking this after August, 2021¹ and no recent version of the database has been

released for the community. Recently, Jaber et al. highlighted the list of ongoing clinical trials of potentially effective drugs for treating COVID-19 in the Gulf Cooperation Council (GCC) countries (9). The authors categorized the drugs in clinical trials into five categories: (a) antiviral (e.g., remdesivir, favipiravir, and darunavir); (b) antiviral repurposed (e.g., hydroxychloroquine, artemisinin, ACE Inhibitors, etc.); (c) anti-inflammatory (e.g., tocilizumab, zafirlukast, anakinra, pioglitazone, etc.); (d) anti-coagulant (e.g., bivalirudin); and (e) miscellaneous (e.g., rivaroxaban, estrogen, iloprost, etc.).

Recently, we developed a knowledgebase, COVID-19Base v2 (10), that highlights COVID-19-related biomedical entities by orchestrating natural language processing techniques, sentiment analysis, and neural networks. To the best of our knowledge, this was the first knowledgebase related to COVID-19 drugs and which also highlighted potential biomedical entities linked to COVID-19 by literature mining. To mine scientific evidence linked to COVID-19, we considered six topic-specific dictionaries (i.e., diseases, Protein Data Bank, drugs, side effects of drugs, genes, and miRNAs). In a co-occurrence-based approach, drug-disease, gene-disease, drug-PDB, and their corresponding sentences from the literature were extracted. The authors used the pre-trained model TextBlob (11), and an unsupervised model based on K-means clustering and the Word2vec model, to determine the sentiment of scores for each disease-drug pair. Subsequently, the authors identified 1,805 diseases, 2,454 drugs, and 1,910 genes associated with coronavirus-related diseases, including COVID-19, through literature mining.

With the above backdrop, it is clear that the scientific community is still trying to find an effective and well-tolerated drug for COVID-19. This ongoing global effort will require the integration of different biomedical entities that are relevant to the infection and the progression of the virus. Therefore, in this paper, we improved our knowledgebase, transitioning it from version 2 to version 3, with the overarching aim of supporting the community in their quest to find a therapeutic treatment for COVID-19. We updated our knowledgebase by adding more entities, i.e., long non-coding RNAs (lncRNAs), alternative medicines (AMs), lncRNA-disease association, and AM-disease associations. Moreover, we considered the recent corpus of literature and clinical trials to provide evidence-based support for the research community.

2. Materials and methods

2.1. Dataset collection

We examined the COVID-19 Open Research Dataset (CORD-19) (12). This dataset covers scholarly articles related to COVID-19 and other diseases related to the coronavirus family (e.g., MERS and SARS). The query “COVID-19” OR “Coronavirus” OR “Corona virus” OR “2019-nCoV” OR “SARS-CoV” OR “MERS-CoV” OR “Severe Acute Respiratory Syndrome OR Middle East Respiratory Syndrome” was used to collect literature from PubMed, PubMed Central, medRxiv, and bioRxiv. CORD-19 has accumulated literature since March 2020 and is still updating the database with relevant literature. For our study, we collected literature that was published after the vaccine rollout, starting from January 2021 up to April 2022, covering 414,899 articles with a

1 <https://chineseantibody.org/opendata/dashboards/covid-19-mab-tracker>

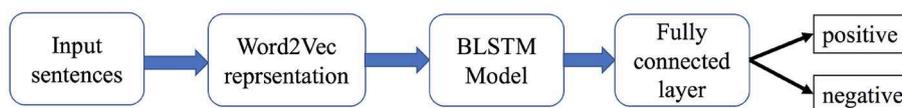


FIGURE 2 Deep-learning model architecture with a Word2Vec embedding layer with BLSTM for drug-disease association.

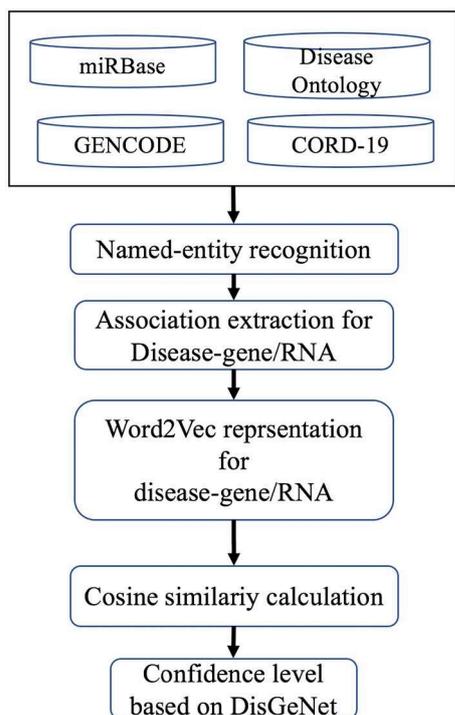


FIGURE 3 Computational workflow for finding associations between disease and genes and RNAs.

maximum, minimum, average, and verified value of DisGeNet scores, respectively. For further details see Khan et al. (10).

2.5. COVID-19Base web server development

For the development of the COVID-19Base v3 web server, we considered a three-tier implementation plan, including a presentation layer, a business layer, and a persistence (database) layer. Additionally, we permitted users to be able to sort and download the query outcome from the web server. The following describes what each layer is responsible for, what technologies they involve, and how they interact with each other (Figure 4).

1. Presentation layer—user interface (UI): represents the front end of the application. It includes static content and dynamic interface elements that users see on the web page. The environment is typically a browser; the technologies

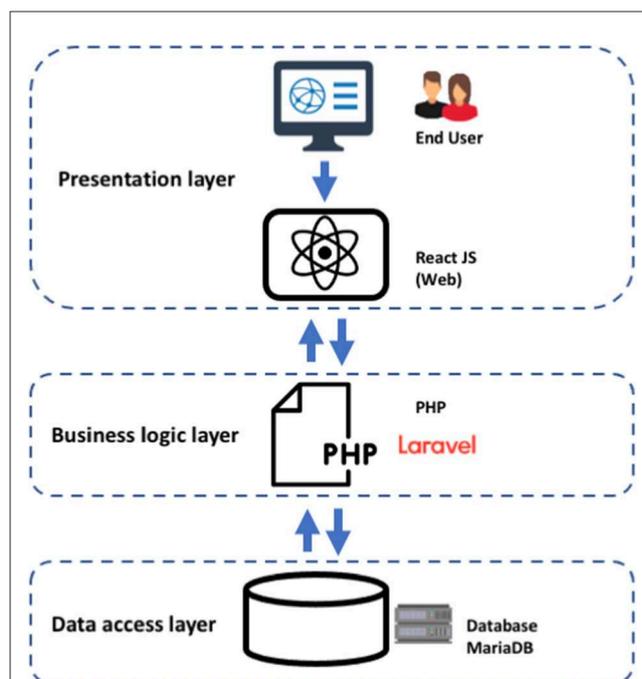


FIGURE 4 Three-tier architecture implementation for the COVID-19Base v3 webserver.

used are HTML, CSS, and JavaScript. We used a React JavaScript framework to build the UI of the COVID-19Base knowledgebase. React is a declarative JavaScript structure for making dynamic client-side applications in HTML. React builds complex interfaces out of simple components, connects them to data on your backend server, and leaves them as HTML. It handles tasteful data-driven interfaces with ease. Additionally, it has all the constructs for building a modern web framework, such as great support for forms, error handling, events, and lists.

2. Business logic or application layer (BLL): part of the application backend contains the app’s business logic and defines internal flows for requests and data. The environments used are a server or infrastructure as a service (IaaS), platform as a service (PaaS), or serverless cloud platforms. The programming languages used for this tier are Java, Python, PHP, JavaScript, or Ruby. We used the Laravel PHP framework to build the COVID-19Base knowledgebase.

3. Persistence or data access layer (DAL): part of the application backend that includes databases and database management systems (DBM) responsible for collecting, managing, and storing information. The environment could be servers, IaaS, PaaS, or serverless cloud platforms. The database management

systems (DBMS) are MySQL, MongoDB, PostgreSQL, MariaDB, Oracle RDBMS, or Redis. We used MariaDB to build the COVID-19Base knowledgebase. MariaDB is an SQL-based database that supports ACID (atomicity, consistency, isolation, and durability)—based data processing for transactions. Atomicity ensures that every transaction is viewed as a single “unit” that totally succeeds or completely fails. If any of the statements that make up a transaction are unsuccessful, the entire transaction is unsuccessful and the database is left untouched. Consistency ensures that a transaction can only bring the database from one valid state to another valid state. Isolation ensures that the concurrent execution of transactions leaves the database in the same state as the sequential execution of the transactions. When a transaction is committed, durability ensures that it will be committed even in the event of a system failure or disaster.

3. Results

3.1. Overall statistics of the biomedical entities in COVID-19Base

Table 1 shows the overall statistics regarding the list of biomedical entities and their association pairs in our knowledgebase. In the current version (v3) of our knowledgebase, we highlighted 4,521 drugs, 2,336 diseases, 2,793 genes, 846 miRNAs, 74 lncRNAs, and 19 AMs. Additionally, we highlighted the association of 96,931 drug-disease pairs, 13,660 drug-PDB pairs, 9,346 disease-gene pairs, 4,458 disease miRNA pairs, and 290

TABLE 1 Summary statistics of the knowledgebase.

Entity	Count(v3)
Drug	4,521
Disease	2,336
Gene	2,793
PDB	950
miRNA	846
lncRNA	74
AM	19
Drug-disease	96,931
Drug-PDB	13,660
Disease-gene	9,346
Disease-miRNA	4,458
Disease-lncRNA	290
COVID-19 and drug	27
COVID-19 and gene	1,315
COVID-19 and miRNA	379
COVID-19 and lncRNA	16

disease-lncRNA pairs. Table 1 also highlights the associations that are related to COVID-19.

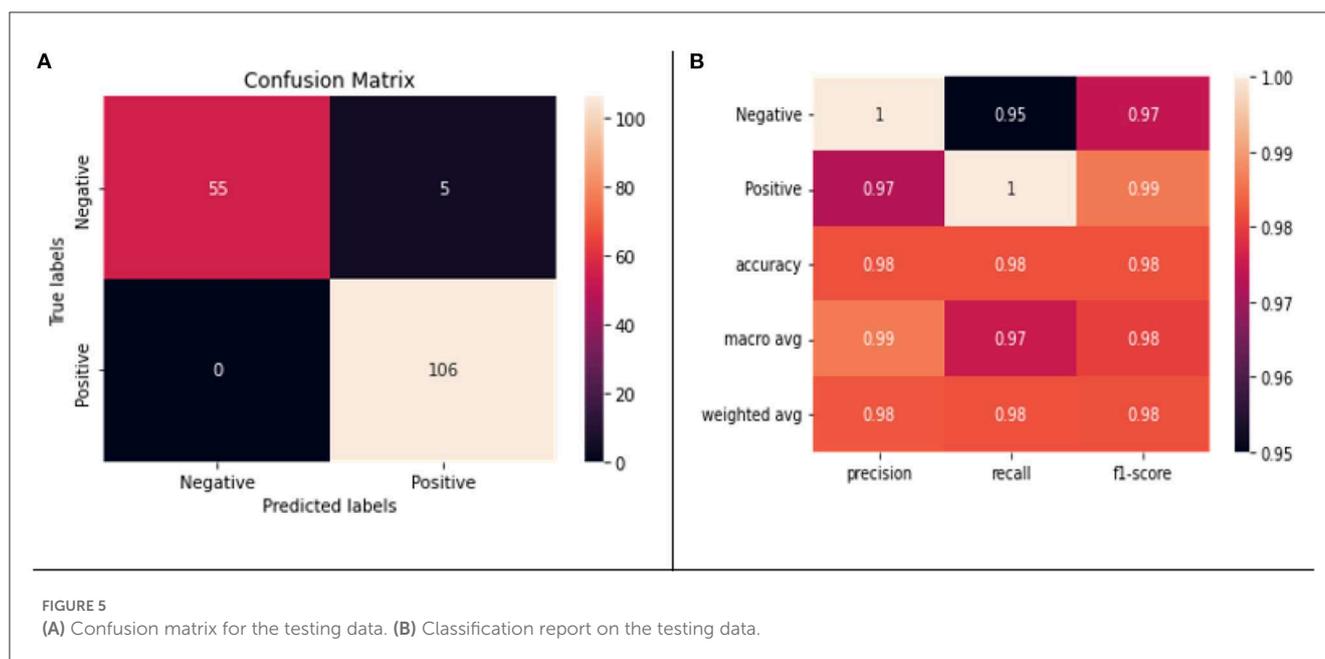
3.2. Performance of the deep-learning-based model for drug-disease associations

We applied fivefold cross validation (CV) to evaluate our model. Figures 5A, B show the confusion matrix and classification report of our model, respectively. There was no misclassification in the positive class; however, five negative samples were classified as positive samples. A weighted accuracy of 97% was achieved by the model, and >90% precision, recall, and F1-scores were achieved for both positive and negative classes.

3.3. Drugs extracted to the knowledgebase

In our knowledgebase, we found 4,521 unique drugs that were associated with different diseases mentioned in the COVID-19 literature. Many of these drugs were associated with COVID-19. Recently the FDA approved Paxlovid for treating COVID-19 patients. Paxlovid is reported to be 89% effective for patients with serious COVID-19 symptoms (24, 25). Ritonavir was found in our knowledgebase, which is packaged with Nirmatrelvir as Paxlovid and is effective with a positive score of 98% in the treatment of COVID-19 patients. But Paxlovid is not free from side effects (26) and extreme care should be taken prior to finalizing the proper dosage with other treatment options. Therefore, in our knowledgebase, we incorporated the list of side effects, collected from the SIDER database, for all the drugs extracted from the literature. Recently, the FDA issued an emergency use authorization for tocilizumab for treating hospitalized adult COVID-19 patients and pediatric patients of 2 years of age or more.³ Additionally, our computational workflow recognized tocilizumab against COVID-19, sarcoidosis, cholangitis, and different types of cancer. Thus, further investigation on this drug is warranted in the near future. Recent studies have also shown that tocilizumab reduces the risk of mechanical ventilation in hospitalized COVID-19 patients (27). Other studies showed that a tocilizumab treatment group had lower mortality compared with the control group (28), and tocilizumab-based treatment reduces the 28–30-day all-cause mortality rate, ICU admissions, superinfections, and mechanical ventilation (MV) (29). In terms of antiviral drugs in our knowledgebase, we found remdesivir and nine other antiviral drugs: favipiravir, ganciclovir, lopinavir, oseltamivir, peramivir, ribavirin, ritonavir, sofosbuvir, and umifenovir. Remdesivir was issued emergency use authorization by the FDA on the 1st May 2020 for hospitalized COVID-19 patients (30). Later, on the 22nd of October 2020, the FDA approved remdesivir for use with COVID-19 adults and pediatric patients requiring hospitalization (30). Additionally, our

³ https://www.uptodate.com/contents/tocilizumab-drug-information?topicRef=127429&source=see_link



knowledgebase classified remdesivir as an effective drug against COVID-19, with a positive score of 95%.

3.4. PDB extracted to the knowledgebase

In our knowledgebase, we incorporated 950 unique PDB entries associated with 1,366 unique drugs extracted from the literature. The SARS-CoV-2 proteins that have been commonly pharmacologically targeted so far to treat COVID-19 include RNA-dependent RNA polymerase (RdRP) and main protease (Mpro). For example, Mpro has been targeted with Paxlovid™, which is a combination of nirmatrelvir and ritonavir, whereas RdRP has been targeted with molnupiravir. Indeed, our analysis was able to capture these and several PDB entries found in the database belonging to these proteins. For example, one of the PDB entries, 5R7Y, represents Mpro in complex with the compound named Z45617795 and is associated with geranylgeranyl diphosphate, atazanavir, danoprevir, remdesivir, selinexor, amino acids, lisinopril, ATP, bafilomycin A1, and defibrotide in the COVID-19Base 3.0 [<https://covidbase-v3.vercel.app/PDBs/5R7Y>]. Similarly, the PDB entry 7BAK represents Mpro in complex with the inhibitor ebisen and is associated with ebisen, urea, trypsin, and formic acid. On the other hand, PDB entry 6WZO, which corresponds to the SARS-CoV-2 nucleocapsid dimerization domain, is associated with ivermectin, camostat, and ebisen. Additionally, we found many PDBs in our knowledgebase, such as 5REA, 6LU7, 6LZE, and 7AGA, that are related to Mpro, highlighting the continuous effort from the research community to understand the Mpro-related mechanism to inhibit the replication and infection of SARS-CoV-2 (31). Overall, our knowledgebase contains a 21,570 unique combinations of drug-PDB records, indicating the huge effort made by the research community to understand the underlying interaction mechanism between drugs and their target proteins.

3.5. Genes extracted to the knowledgebase

In addition to COVID-19-associated PDB entries and drug combinations, we attempted to include human genes that could be linked to other diseases mentioned in the DO database (15). Our analysis identified 2,793 human genes associated with various diseases, including those associated with cancers (such as BRCA1 and BRCA2) and cardiovascular diseases (such as PDE4D). Selecting COVID-19 under the search bar of our knowledgebase will show all genes that have been shown to be associated with this disease. So far, we have found an association of COVID-19 with 1,315 genes. Further, we have classified each such association as “low”, “medium”, and “high” to provide a better understanding of the disease-gene interaction. Additionally, we have included the abstract and the digital object identifier (DOI) of the publication that was used for the analysis. For instance, the gene ABCA3 (ATP-binding cassette subfamily A member 3, which encodes a protein that is involved in the transport of various molecules from the extracellular to the intracellular side) shows a low level of association with COVID-19 (32). The gene EPHB4 (EPH receptor B4, which encodes a protein that is involved in the development of the nervous system and other systems in humans) shows a medium level of association with COVID-19 (33). Similarly, the gene STAT1, which encodes for the STAT1 protein involved in cytokine signaling, also shows a medium level of association with COVID-19 (34).

3.6. MiRNAs and LncRNAs extracted to the knowledgebase

MiRNAs have been considered as molecules with a high potential for discovering drugs for COVID-19 (35). Many miRNAs associated with COVID-19 and other diseases were found in

our knowledgebase. For example, the host miR-122, a liver-specific miRNA (36), has been shown to bind the SARS-CoV-2 genome (37). Additionally, miR-122 serves as a cofactor for the binding of hepatitis C virus (HCV) for its pathogenesis. Therefore, miravirsin anti-HCV RNA-based drugs should be tested against SARS-CoV-2 infection. Our knowledgebase identified miR-122 as being associated with COVID-19, suggesting the usefulness of the knowledgebase for drug repurposing. Another example is miR-21, which was shown to be effective in controlling the steatosis of HCV infection (38), as well as controlling inflammation in COVID-19 patients (39). The knowledgebase also highlighted a medium association between miR-21 and COVID-19, with a score of 0.21. Recently, cobomarsen (MRG-106), an oligonucleotide inhibitor of miR-155, was shown to be effective at controlling cellular proliferation and T-cell lymphoma (40). Additionally, our knowledgebase identified a medium association between miR-155 and COVID-19, with a score of 0.24.

Our knowledgebase summarized 16 different lncRNAs associated with COVID-19, including XIST, the lncRNA well-known for its role in X chromosome inactivation, and MALAT1 and NEAT1, two well-studied lncRNAs linked to COVID-19 (41). These lncRNAs were shown to be associated with COVID-19 with high confidence in our knowledgebase. MALAT1 has been shown to be downregulated in severe COVID-19 patients (42) and upregulated in mild COVID-19 patients (43) in CD4+ T cells. NEAT1 has been shown to be differentially expressed in bronchoalveolar lavage (BAL) cells in severe and mild COVID-19 patients (44). Differences in gender-related lethality for COVID-19 is still under investigation by many research groups (45, 46). Interestingly, the protein coding genes for the ACE2 receptor and the immune regulatory protein TLR7 are located on the X chromosome. Recently, Yu et al. suggested that the dysregulation of XIST is involved in the different immune response between men and women in COVID-19 (47).

3.7. Alternative medicines extracted to the knowledgebase

Alternative medicine (AM) as a treatment for disease is an area that has been relatively untapped by the scientific community. However, considering the urgency of the current pandemic and the shortage of medical facilities, AM is being considered as an adjuvant treatment plan for COVID-19 by people across the world (48). Literature mining identified 19 AMs associated with COVID-19 in our knowledgebase. As suggested by Jaber et al. (20), these AMs can be grouped into the following three categories: (a) vitamins and minerals: vitamins B, C, and D, selenium, and sodium bicarbonate; (b) dietary supplements: lactoferrin, resistant starch, and SivoMixx; and (c) herbal: cannabis/cannabidiol, Chinese preparations, fuzheng huayu, Guduchi Ghan Vati, honey, Kan Jang, Nigella/Nigella Sativa, and resveratrol. Among herbal medicines, cannabidiol, one of the constituents of cannabis, has been shown to be effective at blocking SARS-CoV-2 entry to the cell by controlling ACE2 (49). A clinical trial, NCT03944447⁴, is currently investigating the COVID-19 infection rate in medical cannabis

users against the infection rate in the general population. In terms of dietary supplements, SivoMixx is being clinically trialed NCT⁵ in Italy to evaluate its efficacy in controlling acute diarrhea in COVID-19 patients. Lactoferrin is an iron-binding glycoprotein that has been shown to provide defense against pathogens (50). Multiple clinical trials^{6, 7} are currently in progress to measure the efficacy of lactoferrin against COVID-19. Among minerals, sodium bicarbonate is historically well-known for its use against Spanish Flu in the last century (51). Two clinical trials (NCT04655716⁸ and NCT04530448⁹) are ongoing in the UK and the USA, respectively, to explore its role in the alkalinization of urine to prevent acute kidney injury resulting from progression of COVID-19. The above results indicate that there is an ongoing quest to find an adjuvant treatment plan for COVID-19 that leverages existing AMs. We believe that our knowledgebase will support the proper utilization of AMs during the pandemic based on scientific evidence.

4. Discussion

In the present version (v3) of COVID-19Base, we covered eight biomedical entities and their association with the scientific literature and information from ongoing clinical trials. The previous version (v2) of COVID-19Base was developed based on the literature that was published before the vaccine rollout. In COVID-19Base v2, we identified 2,454 drugs and 1,805 diseases, among other biomedical entities that were mentioned in the literature. During that period, the COVID-19 dataset was covering literature not only for COVID-19 but also for other coronavirus-related diseases, such as SARS and MERS. As a result, the numbers of entities in v2 were higher. However, after the vaccine rollout, researchers were more focused on COVID-19 disease and much of the literature in COVID-19 was focused on the discovery of therapeutic solutions. As a result, the number of unique drugs mentioned in v3 of the COVID-19Base knowledgebase increased significantly. We found 4,521 (as opposed to 2,454 in v2) drugs and 2,336 (as opposed to 1,805 in v2) diseases mentioned in the current version (v3) of COVID-19Base.

In our knowledgebase, we systematically incorporated COVID-19 and other diseases and their association with drugs and the side effects of drugs, PDB entries, protein-coding genes, miRNAs, lncRNAs, and AM. Our knowledgebase incorporates all of these biomedical entities in a single platform, making it convenient for researchers to formulate holistic ideas about all COVID-19-related entities. Moreover, we could not find any such knowledgebase in the literature regarding COVID-19.

Finally, a brief discussion on the methodical advances of COVID-19Base v3 over its predecessor (i.e., COVID-19Base v2) is in order. Among others, a notable difference lies in the use of a more sophisticated deep-learning model architecture

5 <https://clinicaltrials.gov/ct2/show/NCT04368351>

6 <https://clinicaltrials.gov/ct2/show/NCT04412395>

7 <https://clinicaltrials.gov/ct2/show/NCT04427865>

8 <https://clinicaltrials.gov/ct2/show/study/NCT04655716>

9 <https://clinicaltrials.gov/ct2/show/study/NCT04530448>

4 <https://clinicaltrials.gov/ct2/show/NCT03944447>

with a Word2Vec embedding layer with a bidirectional LSTM model for drug-disease association. This is in contrast to v2, in which a relatively simple deep neural network model was used with three features, namely the minimum distance between the disease and the drug terms in the corresponding sentences, the polarity found from the pre-trained TextBlob model, and the sentiment score from an unsupervised model (based on K-means clustering and the Word2Vec model). This sophisticated model in combination with two other factors, namely a more careful manual curation for the ground truth and a more COVID-19-focused dataset (i.e., the latest version of CORD-19), resulted in a huge boost in model performance. It is important to emphasize that in COVID-19Base v3 we provided the sentiment score for each individual sentence related to drug-disease pairs. Therefore, the sentiment score for each individual sentence might look confusing considering our overall knowledge of the efficacy of a drug against COVID-19. For example, in COVID-19Base v3, many sentences related to hydroxychloroquine and COVID-19 pairs were positive as well as negative. Additionally, we know that hydroxychloroquine showed promising results at the early stage of COVID-19, but gradually over time, through multiple research experiments, it was shown to be ineffective against COVID-19 (52). Therefore, users need to consider the sentiment of each sentence in COVID-19Base v3 as a source of evidence for a particular study and should not consider it as a holistic sentiment of a drug-disease pair.

Our study has some limitations that should be addressed. We used literature from the CORD-19 dataset but there are other databases collecting COVID-19-related literature, such as LitCovid (53). Therefore, we may have missed some articles that were collected by those databases. However, overall, the coverage of the CORD-19 dataset is huge and is widely accepted in the community. Moreover, given the limitations of our computational capacity, we focused on literature abstracts, unless otherwise mentioned. However, we believe this also provides much cleaner results than the mining of full literature texts. For AM, many clinical trials have not issued a particular outcome and some of them are suspended, withdrawn, or terminated. Therefore it was difficult to reach any conclusion on the efficacy of AM from the clinical trials registry. Additionally, target structures and their interacting compounds were not linked through the PDB and PubChem databases. In future, we will

consider incorporating this link to make this knowledgebase more informative.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

SAB, RQ, and SM: programming and machine-learning modeling. SAB and SM: website development. TA: conception, design, funding acquisition, and original draft. All authors contributed to the writing of the manuscript and approved the submitted version.

Funding

This work was supported by the Qatar National Research Fund (grant no RRC02-0805-210019 for TA). The open access fee for the article was provided by the Qatar National Library (QNL), Qatar.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Jony SSR, Shahriyar R, Rahman MS, Rahman MS, Alam T. Linking tweets towards geo-localized policies: COVID-19 perspective. *Stud Health Technol Inform.* (2022) 290:709–13. doi: 10.3233/SHTI220170
- Rudan I, Adeloje D, Sheikh A. COVID-19: Vaccines, efficacy and effects on variants. *Curr Opin Pulmon Med.* (2022) 28:180–91. doi: 10.1097/MCP.0000000000000868
- Shah U, Biswas MR, Dolaat KMM, Househ M, Shah Z, Alam T. Vaccine rollout and shift in public sentiment: Twitter-based surveillance study. *Stud Health Technol Inform.* (2022) 290:704–8. doi: 10.3233/SHTI220169
- Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: A mathematical modelling study. *Lancet Infect Dis.* (2022) 22:1293–302. doi: 10.1016/S1473-3099(22)00320-6
- Venkatesan P. Repurposing drugs for treatment of COVID-19. *Lancet Respirat Med.* (2021) 9:e63. doi: 10.1016/s2213-2600(21)00270-8
- Vuorio A, Kovanen PT, Raal F. Cholesterol-lowering drugs for high-risk hypercholesterolemia patients with COVID-19 while on Paxlovid™ therapy. *Future Virol.* (2022) 2022:60. doi: 10.2217/fvl-2022-0060
- Yang L, Liu W, Yu X, Wu M, Reichert JM, Ho M. COVID-19 antibody therapeutics tracker: A global online database of antibody therapeutics for the prevention and treatment of COVID-19. *Antib Ther.* (2020) 3:205–12. doi: 10.1093/abt/tbaa020
- Ullah S, Ullah A, Rahman W, Ullah F, Khan SB, Ahmad G, et al. An innovative user-friendly platform for COVID-19 pandemic databases and resources. *Comput Methods Programs Biomed Update.* (2021) 1:100031. doi: 10.1016/j.cmpbup.2021.100031

9. Jaber BA, Qureshi R, Abd-Alrazaq A, Rahman MA, Househ M, Shah Z, et al. Clinical trials on alternative medicines for COVID-19. *Stud Health Technol Inform.* (2022) 295:366–9. doi: 10.3233/SHTI220740
10. Khan JY, Khondaker MTI, Hoque IT, Al-Absi HRH, Rahman MS, Guler R, et al. Toward preparing a knowledge base to explore potential drugs and biomedical entities related to COVID-19: Automated computational approach. *JMIR Med Inform.* (2020) 8:e21648. doi: 10.2196/21648
11. Dataset. *TextBlob: Simplified Text Processing – TextBlob 0.16.0 Documentation* (2018). Available online at: <https://textblob.readthedocs.io/en/dev/> (accessed January 20, 2023).
12. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. *CARD-19: The COVID-19 Open Research Dataset*. ArXiv Publisher: arXiv. (2020).
13. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: A comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* (2006) 34:D668–72. doi: 10.1093/nar/gkj067
14. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* (2016) 44:D1075–9. doi: 10.1093/nar/gkv1075
15. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Res.* (2019) 47:D955–62. doi: 10.1093/nar/gky1032
16. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.* (2017) 45:D619–25. doi: 10.1093/nar/gkw1033
17. Griffiths-Jones S. miRBase: The MicroRNA sequence database. *MicroRNA Protocols.* (2006) 2006:129–38. doi: 10.1385/1-59745-123-1:129
18. Musleh S, Alazmi M, Alam T. GENCODE annotation for the human and mouse genome: A user perspective. *Practical Guide Life Sci Datab.* (2021) 16:1–25. doi: 10.1007/978-981-16-5812-9_1
19. Piñero J, Bravo I, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* (2017) 45:D833–9. doi: 10.1093/nar/gkw943
20. Jaber BA, Househ M, Shah Z, Rahman MS, Hamdi M, Biswas KH, et al. Gulf cooperation council clinical trials in the pursuit of medications for COVID-19. *Stud Health Technol Inform.* (2022) 289:9–13. doi: 10.3233/SHTI210846
21. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *ArXiv Preprint.* (2013) 2013:13013781. doi: 10.48550/arXiv.1301.3781
22. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv Preprint.* (2015) 2015:150801991. doi: 10.48550/arXiv.1508.01991
23. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA (2015).
24. Mahase E. COVID-19: Pfizer's paxlovid is 89% effective in patients at risk of serious illness, company reports. *Br Med J Publ Gr.* (2021). doi: 10.1136/bmj.n2713
25. Najjar-Debbiny R, Gronich N, Weber G, Khoury J, Amar M, Stein N, et al. Effectiveness of paxlovid in reducing severe COVID-19 and mortality in high risk patients. *Clin Infect Dis.* (2022) 2022:ciac443. doi: 10.1093/cid/ciac443
26. Fishbane S, Hirsch JS, Nair V. Special considerations for paxlovid treatment among transplant recipients with SARS-CoV-2 infection. *Am J Kidney Dis.* (2022) 79. doi: 10.1053/j.ajkd.2022.01.001
27. Tleyjeh IM, Kashour Z, Damlaj M, Riaz M, Tlayjeh H, Altannir M, et al. Efficacy and safety of tocilizumab in COVID-19 patients: A living systematic review and meta-analysis. *Clin Microbiol Infect.* (2021) 27:215–27. doi: 10.1016/j.cmi.2020.10.036
28. Sarfraz A, Sarfraz Z, Sarfraz M, Aftab H, Pervaiz Z. Tocilizumab and COVID-19: A meta-analysis of 2,120 patients with severe disease and implications for clinical trial methodologies. *Turk J Med Sci.* (2021) 51:890–7. doi: 10.3906/sag-2010-131
29. Avni T, Leibovici L, Cohen I, Atamna A, Guz D, Paul M, et al. Tocilizumab in the treatment of COVID-19—A meta-analysis. *QJM.* (2021) 114:577–86. doi: 10.1093/qjmed/hcab142
30. Rubin D, Chan-Tack K, Farley J, Sherwat A. FDA approval of remdesivir—A step in the right direction. *N Engl J Med.* (2020) 383:2598–600. doi: 10.1056/NEJMp2032369
31. Geethakumari AM, Ahmed WS, Rasool S, Fatima A, Nasir Uddin SM, Aouida M, et al. A genetically encoded BRET-based SARS-CoV-2 M protease activity sensor. *Commun Chem.* (2022) 5:117. doi: 10.1038/s42004-022-00731-2
32. Yavuz S, Alsamhoury R, Francis N. Childhood interstitial lung disease masquerading as post COVID-19 respiratory distress. *Cureus.* (2021) 13:e20061. doi: 10.7759/cureus.20061
33. Klaric L, Gisby JS, Papadaki A, Muckian MD, Macdonald-Dunlop E, Zhao JH, et al. Mendelian randomisation identifies alternative splicing of the FAS death receptor as a mediator of severe COVID-19. *MedRxiv.* (2021). doi: 10.1101/2021.04.01.21254789v1.abstract
34. Rincon-Arevalo H, Aue A, Ritter J, Szelenksi F, Khadzhyov D, Zickler D, et al. Altered increase in STAT1 expression and phosphorylation in severe COVID-19. *Eur J Immunol.* (2022) 52:138–48. doi: 10.1002/eji.202149575
35. Ying H, Ebrahimi M, Keivan M, Khoshnam SE, Salahi S, Farzaneh M. miRNAs: A novel strategy for the treatment of COVID-19. *Cell Biol Int.* (2021) 45:2045–53. doi: 10.1002/cbin.11653
36. Alam T, Uludag M, Essack M, Salhi A, Ashoor H, Hanks JB, et al. FARNAs: Knowledgebase of inferred functions of non-coding RNA transcripts. *Nucleic Acids Res.* (2017) 45:2838–48. doi: 10.1093/nar/gkw973
37. Alam T, Lipovich L. miRCOVID-19: Potential targets of human miRNAs in SARS-CoV-2 for RNA-based drug discovery. *Noncoding RNA.* (2021) 7:10018. doi: 10.3390/nrna7010018
38. Clément S, Sobolewski C, Gomes D, Rojas A, Goossens N, Conzelmann S, et al. Activation of the oncogenic miR-21-5p promotes HCV replication and steatosis induced by the viral core 3a protein. *Liver Int.* (2019) 39:1226–36. doi: 10.1111/liv.14112
39. Das D. *Molecular Insights Into the Benefits of Regular Exercise in Combating Immunological Mayhem During SARS-CoV-2 Infection.* (2021). Available online at: <https://www.researchsquare.com/article/rs-845977/latest.pdf> (accessed January 20, 2023).
40. Seto AG, Beatty X, Lynch JM, Hermreck M, Tetzlaff M, Duvic M, et al. Cobomarsen, an oligonucleotide inhibitor of miR-155, co-ordinately regulates multiple survival pathways to reduce cellular proliferation and survival in cutaneous T-cell lymphoma. *Br J Haematol.* (2018) 183:428–44. doi: 10.1111/bjh.15547
41. Rodrigues AC, Adamoski D, Genelhould G, Zhen F, Yamagato GE, Araujo-Souza PS, et al. NEAT1 and MALAT1 are highly expressed in saliva and nasopharyngeal swab samples of COVID-19 patients. *Mol Oral Microbiol.* (2021) 36:291–4. doi: 10.1111/omi.12351
42. Hewitson JP, West KA, James KR, Rani GF, Dey N, Romano A, et al. Suppresses immunity to infection through promoting expression of Maf and IL-10 in Th cells. *J Immunol.* (2020) 204:2949–60. doi: 10.4049/jimmunol.1900940
43. Masoumi F, Ghorbani S, Talebi F, Branton WG, Rajaei S, Power C, et al. Malat1 long noncoding RNA regulates inflammation and leukocyte differentiation in experimental autoimmune encephalomyelitis. *J Neuroimmunol.* (2019) 328:50–9. doi: 10.1016/j.jneuroim.2018.11.013
44. Huang K, Wang C, Vagts C, Raguveer V, Finn PW, Perkins DL. Long non-coding RNAs (lncRNAs) NEAT1 and MALAT1 are differentially expressed in severe COVID-19 patients: An integrated single-cell analysis. *PLoS ONE.* (2022) 17:e0261242. doi: 10.1371/journal.pone.0261242
45. Penna C, Mercurio V, Tocchetti CG, Pagliaro P. Sex-related differences in COVID-19 lethality. *Br J Pharmacol.* (2020) 177:4375–85. doi: 10.1111/bph.15207
46. Li AJ, Li X. Sex-dependent immune response and lethality of COVID-19. *Stem Cell Res.* (2020) 50:102116. doi: 10.1016/j.scr.2020.102116
47. Yu B, Qi Y, Li R, Shi Q, Satpathy AT, Chang HY. B cell-specific XIST complex enforces X-inactivation and restrains atypical B cells. *Cell.* (2021) 184:1790–803.e17. doi: 10.1016/j.cell.2021.02.015
48. Badakhsh M, Dastras M, Sarchahi Z, Doostkami M, Mir A, Bouya S. Complementary and alternative medicine therapies and COVID-19: A systematic review. *Rev Environ Health.* (2021) 36:443–50. doi: 10.1515/revh-2021-0012
49. Esposito G, Pesce M, Seguela L, Sanseverino W, Lu J, Corpetti C, et al. The potential of cannabidiol in the COVID-19 pandemic. *Br J Pharmacol.* (2020) 177. doi: 10.1111/bph.15157
50. Legrand D, Ellass E, Carpentier M, Mazurier J. Lactoferrin: A modulator of immune and inflammatory responses. *Cell Mol Life Sci.* (2005) 62. doi: 10.1007/s00018-005-5370-2
51. Franchini AF, Auxilia F, Galimberti PM, Piga MA, Castaldi S, Porro A. COVID 19 and Spanish flu pandemics: All it changes, nothing changes. *Acta Bio Medica.* (2020) 91:245. doi: 10.23750/abm.v9i12.9625
52. Avezum Á, Oliveira GB, Oliveira H, Lucchetta RC, Pereira VF, Dabarian AL, et al. Hydroxychloroquine versus placebo in the treatment of non-hospitalised patients with COVID-19 (COPE–Coalition V): A double-blind, multicentre, randomised, controlled trial. *Lancet Region Health Americas.* (2022) 11:100243. doi: 10.1016/j.lana.2022.100243
53. Chen Q, Allot A, Lu Z. LitCovid: An open database of COVID-19 literature. *Nucleic Acids Res.* (2021) 49:D1534–40. doi: 10.1093/nar/gkaa952