JMIR Publications
The Leading eHealth Publisher

Towards Preparing a Knowledgebase to Explore Potential Drugs and Biomedical Entities Related to COVID-19

Junaed Younus Khan, Md. Tawkat Islam Khondaker, Iram Tazim Hoque, Hamada R. H. Al-Absi, Mohammad Saifur Rahman, Reto Guler, Tanvir Alam, M. Sohel Rahman

Submitted to: JMIR Medical Informatics
on: June 26, 2020

Table of Contents

Towards Preparing a Knowledgebase to Explore Potential Drugs and Biomedical Entities Related to COVID-19

Junaed Younus Khan[1] BSc; Md. Tawkat Islam Khondaker[1] BSc; Iram Tazim Hoque[1] BSc; Hamada R. H. Al-Absi[2] MSc; Mohammad Saifur Rahman[1] PhD; Reto Guler[3, 4, 5] PhD; Tanvir Alam[2] PhD; M. Sohel Rahman[1] PhD

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

Corresponding Author:
Tanvir Alam PhD
College of Science and Engineering
Hamad Bin Khalifa University
Education City
Doha
QA

Abstract

Background: Novel coronavirus disease 2019 (COVID-19) is taking a huge toll on public health. Along with the non-therapeutic preventive measurements, scientific efforts are currently focused, mainly, on the development of vaccines and pharmacological treatment with existing drugs. Summarizing evidences from scientific literatures on the discovery of treatment plan of COVID-19 under a platform would help the scientific community to explore the opportunities in a systematic fashion.

Objective: The aim of this study is to explore the potential drugs and biomedical entities related to coronavirus related diseases, including COVID-19, that are mentioned on scientific literature through an automated computational approach.

Methods: We mined the information from publicly available scientific literature and related public resources. Six topic-specific dictionaries, including human genes, human miRNAs, diseases, Protein Databank, drugs, and drug side effects, were integrated to mine all scientific evidence related to COVID-19. We employed an automated literature mining and labeling system through a novel approach to measure the effectiveness of drugs against diseases based on natural language processing, sentiment analysis, and deep learning. We also applied the concept of cosine similarity to confidently infer the associations between diseases and genes.

Results: Based on the literature mining, we identified 1805 diseases, 2454 drugs, 1910 genes that are related to coronavirus related diseases including COVID-19. Integrating the extracted information, we developed the first knowledgebase platform dedicated to COVID-19, which highlights potential list of drugs and related biomedical entities. For COVID-19, we have highlighted multiple case studies on existing drugs along with a confidence score for their applicability in the treatment plan. The resulting knowledgebase is made available as an open source tool, named COVID-19Base, for the scientific community: http://77.68.43.135:97/search/.

Conclusions: Proper investigation of the mined biomedical entities along with the identified interactions among those would help the research community to discover possible ways for the therapeutic treatment of COVID-19.

(JMIR Preprints 26/06/2020:21648)
DOI: https://doi.org/10.2196/preprints.21648

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Towards Preparing a Knowledgebase to Explore Potential Drugs and Biomedical Entities Related to COVID-19

Junaed Younus Khan[1], Md. Tawkat Islam Khondaker[1], Iram Tazim Hoque[1], Hamada R. H. Al-Absi[2], Mohammad Saifur Rahman[1], Reto Guler[3,4,5], Tanvir Alam[2,*], M. Sohel Rahman[1]

[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

[2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

[3]International Centre for Genetic Engineering and Biotechnology (ICGEB), Cape Town-Component, Cape Town, South Africa.
[4]Institute of Infectious Diseases and Molecular Medicine (IDM), Department of Pathology, Division of Immunology and South African Medical Research Council (SAMRC) Immunology of Infectious Diseases, Faculty of Health Sciences, University of Cape Town, South Africa.
[5]Wellcome Centre for Infectious Diseases Research in Africa, Institute of Infectious Disease and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town, South Africa.
*talam@hbku.edu.qa : To whom correspondence should be addressed.

Abstract

Background: Novel coronavirus disease 2019 (COVID-19) is taking a huge toll on public health. Along with the non-therapeutic preventive measurements, scientific efforts are currently focused, mainly, on the development of vaccines and pharmacological treatment with existing drugs. Summarizing evidences from scientific literatures on the discovery of treatment plan of COVID-19 under a platform would help the scientific community to explore the opportunities in a systematic fashion.

Objective: The aim of this study is to explore the potential drugs and biomedical entities related to coronavirus related diseases, including COVID-19, that are mentioned on scientific literature through an automated computational approach.

Methods: We mined the information from publicly available scientific literature and related public resources. Six topic-specific dictionaries, including human genes, human miRNAs, diseases, Protein Databank, drugs, and drug side effects, were integrated to mine all scientific evidence related to COVID-19. We employed an automated literature mining and labeling system through a novel approach to measure the effectiveness of drugs against diseases based on natural language processing, sentiment analysis, and deep learning. We also applied the concept of cosine similarity to confidently infer the associations between diseases and genes.

Results: Based on the literature mining, we identified 1805 diseases, 2454 drugs, 1910 genes that are related to coronavirus related diseases including COVID-19. Integrating the extracted information, we developed the first knowledgebase platform dedicated to COVID-19, which highlights potential list of drugs and related biomedical entities. For COVID-19, we have highlighted multiple case studies on existing drugs along with a confidence score for their applicability in the treatment plan. The resulting knowledgebase is made available as an open source tool, named COVID-19Base, for the scientific community:

http://77.68.43.135:97/search/.

Conclusions: Proper investigation of the mined biomedical entities along with the identified interactions among those would help the research community to discover possible ways for the therapeutic treatment of COVID-19.

Keywords: COVID-19; 2019-nCoV; Coronavirus; SARS-CoV-2; SARS;

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) initially spread widely in China, then in Italy, and, now in other parts of the world, causing coronavirus disease 2019 (COVID-19) [1, 2]. SARS-CoV-2 is the novel coronavirus that causes the alarming pandemic of COVID-19 [3]. While the novel coronaviruses (SARS-CoV-2) have gained enormous attention and grave concern as a consequence of the current worldwide COVID-19 pandemic situation, other known human coronaviruses, beta coronaviruses (SARS-CoV, MERS, OC43, HKU1) and alpha coronaviruses (229E, NL63), historically, caused a significant level of public-health concerns and resulted in severe respiratory syndrome for the patients as well [4]. To combat the global lethality of COVID-19, the public demands an urgent solution for the detection and therapeutic treatment, which requires a comprehensive experimental elucidation of relevant biomedical entities (e.g., genes, non-coding RNAs (ncRNA), viruses, drugs, etc.) [5]. But this is a relatively slow process due to the inherent nature of the experimental validation. As an alternative, faster *in silico* methods are applied [6, 7] which can act at least as a filter before actual wet lab validation. Virtual screening, molecular docking and other *in silico* methods are already been investigated to discover drugs against COVID-19 [8]. Still, it is also a daunting task due to a large number of possible combinations of biomedical entities (e.g., drug-gene pairs) that need to be examined [9]. To mediate these two extreme experimental setups, and at the same time, to enable a comprehensive exploration of potential therapeutic treatments, knowledgebase based solutions are proposed, as part of this study, for the scientific community to focus on a relatively smaller number of potential biomedical entities that may guide to the discovery of a novel solution of the COVID-19 treatment.

In line with the spirit alluded above, there exist databases that focused on the virus-related diseases for multiple hosts. For example, in ViRBase [10], the authors highlighted the association between non-coding RNAs (ncRNAs) and viruses in 20 hosts. VISDB database, based on literature mining, integrated the virus interaction site in humans for five DNA oncoviruses and four RNA retroviruses [11]. The Virus Pathogen Resources (VIPR), developed a portal collecting a comprehensive set of information related to coronavirus and hepatitis C virus (HCV), and other viruses [12, 13]. But none of the mentioned databases are much useful for COVID-19/SARS-CoV-2, as those databases were not specific to novel coronavirus, or they provide very limited information about the associated genes, or do not cover other associated factors that are involved in coronavirus-related diseases, drugs and their side effects. Moreover, there exists no such knowledgebase that has integrated all biomedical entities specific to COVID-19/SARS-CoV-2. To fill this gap in a timely manner, we explore the potential of machine intelligence to automatically mine the scientific literature with a goal to develop the first comprehensive knowledgebase that integrates several biomedical entities associated with COVID-19/SARS-CoV-2. To achieve this, we have leveraged the state-of-the-art natural language processing algorithms, sentiment analysis, and deep learning-based techniques and applied those on the largest corpus of coronavirus related scientific literature.

Materials and Methods

Datasets

For this study, we considered COVID-19 Open Research Dataset (CORD-19) [14], generated by Allen Institute of AI. The dataset contains over 138K scholarly articles, related to COVID-19 and the coronavirus family of viruses. The dataset was collected using 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” against PubMed, PubMed Central (PMC), bioRxiv and medRxiv pre-prints and it covers major part of all research articles related to COVID-19 and other coronavirus (e.g., MERS, SARS, etc.) till 09-06-2020. Unless otherwise specified, we considered both the abstracts and full body (when available) from the research articles for downstream analysis.

Source of dictionaries

We collected the gene names from HGNC [15], PDB entries from PDB [16], miRNA from miRBase [17], disease names from DO [18], drug names from DrugBank [19], and the drug side effects from SIDER [20].

Overview of the Methodology

We extracted disease-drug, disease-gene, drug-PDB pairs, and their corresponding sentences from the CORD-19 literature in a co-occurrence based approach. For evaluating the effectiveness of the disease-drug pairs, we used both pre-trained model (TextBlob) and unsupervised model (developed by us using Word2Vec model and K-means clustering) to determine sentiment scores of the sentences extracted for each pair. We further used these sentiment scores along with the minimum distance between the disease and drug term in the corresponding sentences as input features of our neural network model which we used for the final classification of the disease-drug pairs (as positive or negative). For determining the confidence level of the extracted disease-gene associations, we transformed each disease and gene of a pair into two separate vectors using Word2Vec model and calculated their cosine similarity. We used the known disease-gene associations from DisGeNET database as the gold standard to determine the confidence level of the new associations on the basis of cosine similarity measures. Finally, we extracted the side-effects of the drugs that were found in our mining from SIDER. Additionally, a feedback mechanism has been incorporated into the COVID-19Base to collect feedback from respected users for future use.

Extracting Disease-Drug Interactions

We extracted disease-drug interactions from the CORD-19 literature and classified them in two categories (labels): Positive and Negative. The positive label means the drug is potentially effective for curing the disease, and the negative label means the opposite. We also determined a confidence score which indicates our level of confidence for that automatic label. Fig. 1 shows the workflow of extracting disease-drug interactions and predicting the effectiveness of drugs against diseases with confidence scores.

Disease and Drug Names Extraction

To extract relevant disease-drug pairs from the CORD-19 literature, we employed a dictionary-based approach to detect mentions of diseases and drugs in the literature. We used Disease

Ontology [18] and DrugBank [19] to prepare the disease and drug dictionaries. We leveraged the Aho-Corasick algorithm [21] to search the drug and disease names considering the large size of drug and disease dictionaries and the corpus itself. The Aho-Corasick algorithm is a string searching algorithm that efficiently locates multiple patterns in a large blob of text. The time complexity of the algorithm is $O(n + m + z)$, where n is the length of the text, m is the total length of all the patterns to be searched, and z is the total number of occurrences of the patterns in the text.

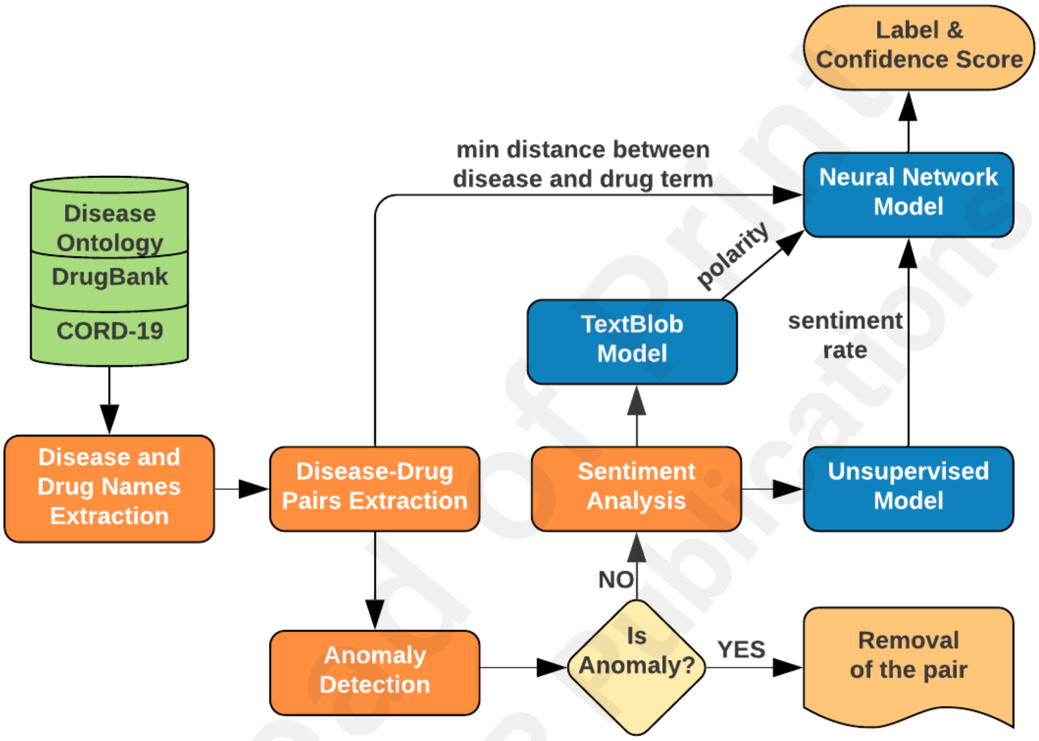


Figure 1. Flowchart of extracting disease-drug interactions and predicting the effectiveness of drugs against diseases with confidence scores.

Disease-Drug Pairs Extraction

After extracting the disease and drug names separately, we wanted to mine the literature and identify the sentences that contain the disease and drug pairs to semantically evaluate their interactions. For this purpose, we searched for every disease-drug pair from our disease and drug list in the CORD-19 literature and collected every sentence where a co-occurrence is found. Then we created a document for every disease-drug pair combining all extracted sentences. Thus, we built a disease-drug pair to document mapping. We did not use any pattern-based approach here as done in the paper [22] as this could result in missing a good number of sentences containing disease-drug pairs.

Anomaly Removal

As we automatically extracted the sentences containing the disease-drug pairs, there was a possibility of aberration in our extracted data. So we decided to check and remove any abnormality from our collected data before going into the next stage of the pipeline. We used unsupervised anomaly detection [23] for doing this job. Unsupervised anomaly detection technique detects anomalies in an unlabeled dataset by looking for instances that seem to fit the least to the remainder of the dataset under the assumption that the majority of the instances in the dataset are ‘normal’. We used K-Means clustering algorithm [24] as it has been used for anomaly detection in several studies [25-29]. We proceed as follows. First, we used Doc2Vec [30] to create a numeric representation of each document associated with each disease-drug pair. Then we fitted these representations into our K-means model and observed two clear clusters of easily discriminable sizes, where the smaller one consisted of only 189 instances. As we know that anomalies differ from the normal instances significantly and occur very rarely in the data, we could assume that the instances of the smaller cluster were indeed anomalies. We also checked a number of instances manually to verify our assumption. We discarded these 189 instances from any further consideration.

Sentiment Analysis

We applied sentiment analysis to automatically assess the effectiveness of a drug to treat a particular disease in the context of each extracted drug-disease pair. First, we applied the concept of transfer learning. We used TextBlob [31] which is a pre-trained sentiment analysis tool provided as a Python library. However, it showed some inconsistency in some cases as expected from a pre-trained model and we felt the necessity of unsupervised sentiment analysis which is the second model of our pipeline. We got a polarity score from the TextBlob model and a sentiment rate from our unsupervised model for each disease-drug pair which were subsequently fed to our Neural Network model to predict the final label.

TextBlob model

TextBlob is a Python library that is widely used in natural language processing tasks such as POS tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Given the sentences that we mined for each disease-drug pair as input, TextBlob gives a polarity score in the range between -1 and 1. We recorded the polarity scores for each disease-drug to use it as a feature for our Neural Network model.

Unsupervised model

We used the concept of K-means clustering again for unsupervised sentiment analysis. First, we trained the Word2Vec [32] model with our mined literature and got a vector representation of every word. Then we ran K-means clustering on the estimated word vectors and found two clusters (positive and negative). The positive cluster was decided on the basis of the presence of several positive words (in the context of a disease-drug pair) in it such as ‘cure’, ‘preclude’, ‘inhibit’, ‘prescribe’, ‘reduce’, ‘modest’, etc. On the other hand, the negative cluster contained words like ‘risky’, ‘kill’, ‘danger’, etc. Then we assigned each word a sentiment value, either +1 or -1, based on the cluster (positive or negative) they belong to. We weighed this value by dividing it by the distance between the word and the centroid of its cluster to describe the extent of its potential positive or negative-ness. Then we calculated the term frequency-inverse document frequency (tf-idf) score [33, 34] of each word in the sentence collection to consider the significance of the unique words. Next, we built a tf-idf representation, T, for each disease-drug pair by replacing each word of the corresponding sentences with its tf-idf score and a sentiment value representation, S, by replacing each word with its sentiment value. Finally, we took their dot product ($T \bigcirc S$) as the final sentiment rate of our unsupervised model.

Neural Network Model for Automatic Labelling and Confidence Score

We used a Deep Neural Network (DNN) model to automatically predict the label and confidence score for our disease-drug pairs. We considered relatively simpler neural network having two hidden layers as such models commonly perform well for smaller dataset compared to other neural networks with many layers and parameters [35, 36].

Training data

We manually labeled 200 disease-drug pairs to train our Neural Network model. Among them, there were 110 positive instances and the rest were negative.

Input features

We used the polarity or sentiment score given by the TextBlob and unsupervised models as the input features for our Neural Network model along with the minimum distance between the disease and drug term in the corresponding document.

Model setup and output

The DNN structure used in this study is similar to that shown in Fig. 2. It consists of 1 input layer with 3 neurons (each neuron corresponds to one input feature), two hidden layers with eight and four neurons respectively, and one output layer containing one neuron for binary classification (positive or negative). The transfer functions of the first and second hidden layers were Rectified Linear Unit (ReLU) [37] and Hyperbolic Tangent function (tanh) [38] respectively. The transfer function of the output layer was a Sigmoid function [39]. We trained the DNN model using Xavier initialization [40] which tries to make the variance of the outputs of a layer to be equal to the variance of its inputs. We used Adam optimizer [41] and the maximum training epoch was set to 500. We split our labeled data into training and test sets on an 80:20 ratio. We trained our model on the training data and achieved 75% accuracy on the test set.

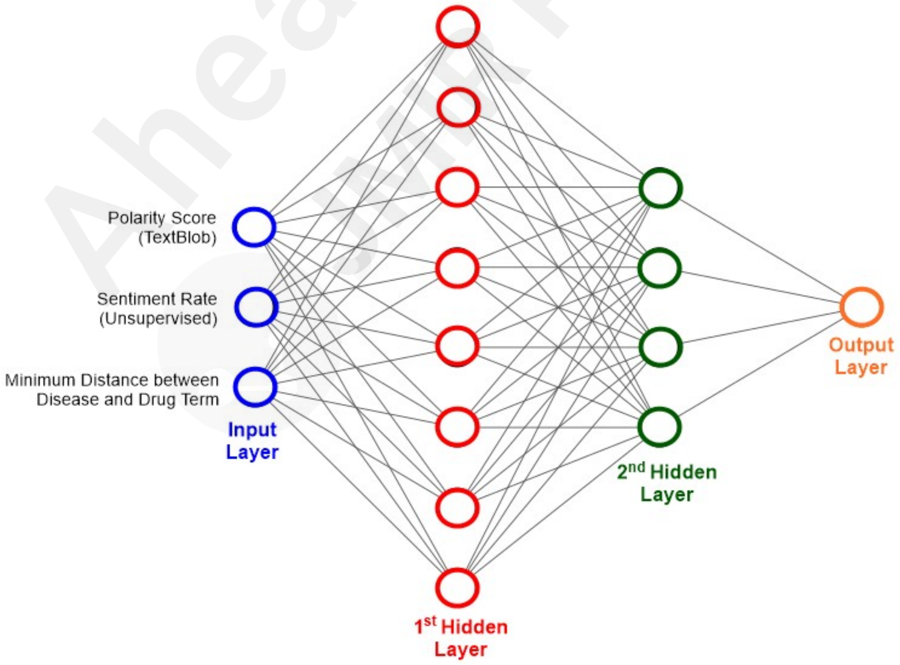

Polarity Score (TextBlob)
Sentiment Rate (Unsupervised)
Minimum Distance between Disease and Drug Term
Input Layer
1st Hidden Layer
2nd Hidden Layer
Output Layer

Figure 2. Schematic diagram of the deep neural network used to predict the effectiveness of drugs against diseases.

Extracting Disease-Gene Associations

Fig. 3 shows the workflow of extracting disease-gene associations. We extracted gene names along with miRNAs from the CORD-19 literature in a dictionary-based approach using HGNC [15] and miRBase [17]. Then we extracted their associations with diseases in a similar process that we had used to extract the disease-drug pairs and collected all the abstracts where a co-occurrence was found. Next, we applied the concept of cosine similarity [42] to confidently infer the associations. We transformed each disease into vector V_1, each gene (and miRNA) into vector V_2, and then calculated the cosine similarity of V_1 and V_2 for each pair. To create the vector representations, we trained a Word2Vec model with all the collected abstracts. We used the DisGeNET [43] database as the gold standard to evaluate the performance of cosine similarity in predicting the gene-disease linkage. First, we calculated the maximum, average, and minimum cosine similarity of the pairs that are common both in our findings and in the DisGeNET database. We found that 99.7% of the newly discovered pairs lie within this range (determined from DisGeNET) in terms of cosine similarity. We further classified the associations in three classes (high, medium, and low) in terms of confidence as follows: pairs having cosine similarity closest to the maximum (minimum) of the known ones were considered as high (low) confidence associations, and the remaining ones (closest to the average) as medium confidence associations. Moreover, pairs that were also found in the DisGeNET database were labeled as verified associations.

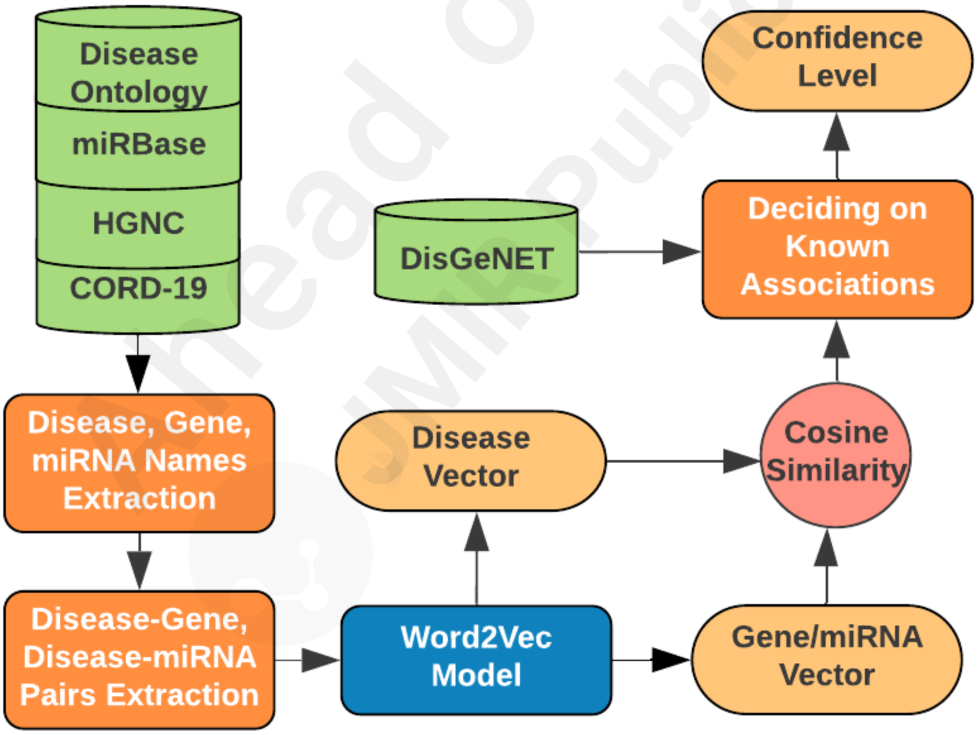


Figure 3. Flowchart of extracting disease-gene, disease-miRNA associations and determining their confidence levels.

Extracting Drug-Protein Associations

We also extracted drug-protein associations from the CORD-19 literature applying the same co-occurrence based approach as mentioned above. We used PDB IDs from the Protein Data Bank [16] for extracting protein names. Unlike the disease-gene associations, we did not apply the concept of cosine similarity here as we did not find any suitable dataset that could be used as the gold standard in this case.

Extracting Side-effects of Drugs

The drugs we are suggesting through this literature mining may come with different side-effects. Therefore, we also explored the possible side-effects of the drugs. We collected the drugs with the corresponding side-effects from SIDER [20] and mapped them with the drugs mentioned in the CORD-19 literature to extract the possible side-effects.

Feedback Mechanism

We implemented a feedback mechanism in the COVID-19Base for future improvement. This mechanism enables the expert users (from the scientific community) to share their valuable feedbacks on the label (positive or negative) for a particular interaction determined by the automatic NLP based approach. The users can voluntarily label each sentence that is mined from the literature as a source of an interaction. These feedbacks will be recorded and further processed to enrich the labeled dataset which can be leveraged in the next version of COVID-19Base to further improve the prediction quality for determining effective disease-drug interactions. The accompanying tutorial (user manual) on COVID-19Base highlighted an example of how a user can use the feedback mechanism.

Results

Terms and interactions thereof highlighted in CORD-19 dataset

Based on our computational workflow, we identified 1805 diseases, 2454 drugs, 1910 genes, 11 miRNAs and 70 PDB entries from the CORD-19 literature (Table 1). Among the disease-drug pairs, 21581 are found to be positive and 1318, negative. Among the disease-gene associations, 2088 are verified (V), 82 associations are found with high-confidence (H), 12231 with medium-confidence (M), and 1488 with low-confidence (L). More results are available in Table 1. Notably, a tiny part (1.5%) of the findings have been manually labeled. Interestingly, we found 194 drug-PDB pairs for coronavirus related diseases which indicates the rapid growth of experimental validation to understand the interaction mechanism of drugs and target proteins.

Table 1. Pairs of terms as identified in the analyzed set of documents. Here V, H, M and L means Verified, High-, Medium- and Low-confidence associations respectively. +ve (-ve) indicates an (not) effective association.

Interaction/Association	# of extracted pairs of terms
Disease - Drug	22899 (21581+ve, 1318-ve)
Disease - Gene	15889(2088V, 82H, 12231M, 1488L)
Disease - miRNA	56 (48 M, 8 L)

Drug - PDB	194

COVID-19 related terms and interactions thereof

Our computational workflow identified 514 drugs and 417 genes that are directly associated to COVID-19 (Table 2). Among the 514 drugs, 492 are found to be positive and 22, negative. Among the 417 genes, 347 are found with medium-confidence (M), and 70 with low-confidence (L).

Table 2. Biomedical terms that are related to COVID-19. Here V, H, M and L means Verified, High-, Medium- and Low-confidence associations respectively. +ve (-ve) indicates an (not) effective association.

Interaction/Association	# of extracted pairs of terms
COVID-19 - Drug	514 (492+ve, 22-ve)
COVID-19 - Gene	417 (347 M,70 L)
COVID-19 - miRNA	3 (2 M, 1 L)

Genes related to COVID-19

Our automated workflow identified C-reactive protein (CRP) as one of the COVID-19 associated genes with “Medium” confidence. CRP is a known clinical biomarkers for SARS [44] and the level of CRP increases significantly in SARS infected patients. The level of CRP was also higher for COVID-19 patients in some clinical cases [45, 46]. More than 25 papers (from the CORD-19 dataset) related to the association between CRP and COVID-19 have been identified through our computational workflow. Furthermore, ELANE, AZU1, MPO, PRTN3, CTSG, TCN1- all these genes were shown to be significantly altered in COVID-19 patients [47], and our automatically prepared knowledgebase highlights all of them to be associated to COVID-19 with “Medium” or “Low” confidence as well. ACE2 and TMPRSS2 genes are known to be involved in SARS-CoV-2 infection [48]; in fact SARS-CoV-2 uses angiotensin-converting enzyme 2 (ACE2) as a receptor for entry into host cells [49, 50]. Spike(S)-protein from SARS-CoV-2 binds with the ACE2 receptor and protease TMPRSS2 mediate the infection process [51]. It is important to note that ACE2 and TMPRSS2 were not directly listed as genes in DisGeNet as associated to COVID-19. In spite of that, our data driven approach based on gold standard dataset from DisGeNet was able to infer the association of ACE2 and TMPRSS2 with COVID-19 with “Medium” confidence, which suggests the efficacy of our approach. Analyzing the complete ACE2 interaction network Wicik et al. [52] listed element genes (ACE2, ANPEP, DPP4, CCL2, MEPIA, TFRC, ADAM17, NPC1, FABP2, TMPRSS2, CLEC4M) and all of these genes are identified as COVID-19 associated in our automatically prepared knowledgebase. We mined three miRNAs (hsa-miR-4661-3p, hsa-miR-429, and hsa-miR-183) that are mentioned in the abstract of COVID-19 related literature.

Case studies

In the rest of this section, we discuss interesting and useful findings from our automatically prepared knowledgebase in the context of potential drugs that can be investigated for the potential therapeutic treatment for COVID-19.

Case study 1: Dexamethasone can be considered as an effective drug for COVID-19

Dexamethasone, an inexpensive and commonly used steroid, is a major breakthrough in the fight against COVID-19. We find Dexamethasone as a positive (i.e., effective) drug for COVID-19, automatically labeled as such through our computational workflow with a confidence score of 77.61%. Our computational workflow also discovers the effectiveness of this drug against Pneumonia, Respiratory failure, and Diarrhea which are strongly correlated to COVID-19 [53, 54]. Thus further exploration of this drug to fight COVID-19 is likely to be fruitful. Recent studies suggest that Dexamethasone reduces the risk of death from 40% to 28% (for patients on ventilators) and from 25% to 20% (for patients needing oxygen) [55].

Case study 2: Ivermectin might be considered as an effective drug for COVID-19

Ivermectin is an effective drug against Pneumonia, Diarrhea, which has recently been claimed to have astounding success in treating patients suffering from COVID-19 as well [56]. It is an FDA-approved drug used for parasitic infections and therefore has a potential for repurposing. It is found that Ivermectin inhibits the replication of SARS-CoV-2 *in vitro* [57]. Recently, a team of medical doctors in Bangladesh has reported quick recoveries of COVID-19 patients using this drug [58]. We find Ivermectin as a positive (i.e., effective) drug for COVID-19, automatically labeled with a confidence score of 77.91%. It is also found to be a positive drug for Pneumonia, Diarrhea in our knowledgebase.

Case study 3: Remdesivir seems effective for COVID-19

Remdesivir has been identified as a positive (i.e., effective) drug for COVID-19, automatically labeled as such through our pipeline with a confidence score of 68.18%. Thus, it seems to be a promising drug for further investigation for treating COVID-19. Interestingly, it is recently being considered as an effective drug for treating COVID-19 [59]. Notably, Remdesivir is an antiviral drug originally developed for Ebola treatment [60, 61]. A recent clinical trial conducted by the National Institute of Allergy and Infectious Diseases (NIAID) shows that Remdesivir helps COVID-19 patients recover faster and improves their survival rates. Adult patients, treated with Remdesivir, were found to recover four days faster, an improvement of 31% compared to other patients and the overall death rate dropped from 11.6% to 8% [62]. Remdesivir is now under consideration of more than ten clinical trials for COVID-19 [63]. We found 6LU7 as one of the PDB entries for Remdesivir and exploring the corresponding literature [64] we found that Remdesivir is shown to be an effective inhibitor for SARS-CoV-2 main protease using molecular docking [65, 66].

Case study 4: Hydroxychloroquine is not an effective treatment plan for COVID-19

Anti-malaria drug Hydroxychloroquine, which is one of the most talked-about drugs for treating COVID-19, has also been found in our mining, albeit with a negative interaction. Our model finds it as a negative (i.e., ineffective) drug with 64.67% confidence. Additionally, it is also revealed that this drug has 111 side-effects including Anaemia, Haemorrhage, Liver disorder, Hepatitis fulminant, Cardiomyopathy, Cardiac failure, etc., which makes it a risky option especially for patients with heart and liver complications. Informatively, although the US Food and Drugs Administration (FDA) had previously granted authorization to use this drug for COVID-19, it recently has cautioned against its' use outside of the hospital setting or a clinical trial due to its side effects and risk factors [67].

Case Study 5: Statin drugs could be effective against COVID-19

Statins are effective as lipid-lowering drugs and mainly used for the treatment for cardiovascular diseases [68]. Statins are also have well known for their anti-inflammatory effects [69], and some studies have supported the rationale of the use thereof as part of COVID-19 treatment protocol [70]. Multiple clinical trials (NCT04343001, NCT04380402) have been launched so far to check the efficacy of Satins against COVID-19 [71, 72]. In our knowledgebase, majority of the statin classes were shown to be effective against COVID-19. For example, Ulinastatin, Rosuvastatin, Fluvastatin, Lovastatin were labeled as positive (i.e., effective) drug against COVID-19 with 94.04%, 79.38%, 78.88%, 70.75% confidence score, respectively. Through our automated computational workflow, we found only one mention of Atorvastatin in the literature [73]. In that single article, Deliwala et al. mentioned Atorvastatin as part of a prevention plan against cortical stroke for a 31-year old COVID-19 female patient without referring to the effectiveness of Atorvastatin against COVID-19. Consequently, our knowledgebase labeled Atorvastatin with negative sentiment with a rather low confidence score (61.22%) against COVID-19. We anticipate that as the number of articles related to Atorvastatin in COVID-19 treatment protocol would improve, the sentiment (effective vs. ineffective) evaluation for this drug would be warranted in the near future. Based on our finding, it is safe to state that statins, as a low-cost and well-tolerated drugs, deserves to be investigated in more detail for clinical trials that may help even the low-middle-income countries where expensive drugs might not be affordable for mass people in this pandemic.

Discussion

Principal Findings

In our knowledgebase, through computational workflow, we not only extracted the drugs and other biomedical terms that are mentioned in the literature, but we also identified “term pairs” based on their co-occurrence, which will allow the scientific community to investigate in depth association between term pairs like disease-gene, disease-drug. A good number of drugs are associated with COVID-19 representing the cumulative effort from the scientific community focusing on drug repurposing than novel drug discoveries; which is a rational approach in a pandemic situation [74]. We leverage an automated approach to highlight the effectiveness of drug against disease based on sentiment analysis of the text mentioned in the literature. We found Dexamethasone, Ivermectin, Remdesivir etc. in the list of potential drugs for the COVID-19 treatment through this literature mining. We highlighted Hydroxychlorquine as an ineffective drug against COVID-19. We extracted the disease gene associations from the literature and based on cosine similarity against gold standard DisGeNet dataset, we provided a confidence level for the associations between diseases and genes. We found 194 drug-PDB associations which show the enormous amount of effort form the scientific community to understand the mechanism behind drug-target interaction and virus:host protein interaction mechanisms for coronavirus related diseases. Surprisingly, we found quite a few numbers of miRNAs related to COVID-19, indicating the primary focus of the scientific community towards protein-based drugs rather than RNA-based drugs, though there have been successful precedents of RNA-based drugs as antiviral agents. One of the successful RNA based precedents is Miravirsen, which binds miR-122 to prevents miR-122 from hybridizing to the hepatitis C virus (HCV) RNA genome, hence depriving HCV of its essential cellular co-factor and blocking HCV replication [75]. We expect more literature along this line in coming months as a potential treatment plan for COVID-19.

Research Implications

Currently we are facing the largest public health emergency since the 1918 influenza outbreak [76]. From the very beginning of this outbreak, the scientific community has invested enormous amount of efforts to find vaccines and therapeutic solutions. Vaccines for SARS-CoV-2 might come too late to have any effect on the first wave of COVID-19 pandemic [77]. However, vaccines might be useful in the subsequent waves of novel coronavirus or in a post-pandemic scenario where the novel coronavirus may continue emerging as a seasonal virus [77]. In this scenario, identification of drugs with a good efficacy and minimal side-effect, would be a rational goal that can be achieved in near future to combat SARS-CoV-2 [48]. Though promising pharmacological results on drug repurposing are emerging every day, unfortunately, no drug has been approved, so far, for the treatment of COVID-19 as of now. Repurposing of the existing drugs, many in preclinical and clinical stages, are under investigation across the world [78]. With the growing information of the SARS-CoV-2 along with the publications on the similar respiratory diseases (e.g., pneumonia, SARS), it would be essential to investigate existing drugs that are already known to be effective in other respiratory diseases. As a prime example, Dexamethasone, an FDA approved drug, was known to be effective against pneumonia [79], respiratory failure [80] and other diseases. But there was no evidence of its effectiveness against COVID-19 until its recent breakthrough in the clinical trial [55]. Though the final approval of the drug is still pending, had it been investigated earlier, more lives could be saved in this pandemic situation.

In line with the above spirit, our effort is expected to support the scientific community and the decision makers to identify candidate drugs with proper evidences from the scientific literature. This will also help the stakeholders to explore the existing drugs that are already known to be effective in other respiratory diseases. While careful manual curation of the identified associations of biomedical entities is the ultimate goal, our novel approach estimates the effectiveness of drug for coronavirus related diseases based on natural language processing, sentiment analysis, and deep learning to support the community to shorten the potential list of drugs, ultimately saving a huge amount of time in this research direction.

Tool and Availability

We have made our computational workflow and the resulting database as an open source tool named COVID-19Base, for the scientific community. COVID-19Base is available at: http://77.68.43.135:97/search/. It does not only identify the terms and associations; it highlights the relevant literature through the digital object identifier (DOI) information so that any expert/scientist exploring this tool can check the main source him/herself for more detailed information. As the number of scientific publications, particularly, on COVID-19 is surging, we will update the knowledgebase on a monthly basis and integrate all the recent updates in the knowledgebase. COVID-19Base has already gone through its first transformation, COVID-19Base 1.0 to COVID-19Base 2.0 as during the manuscript preparation phase the CORD-19 dataset was updated. Informatively, earlier version of the CORD-19 dataset contained only around 44K papers whereas the current version covers more than 138K scholarly articles. The knowledgebase materials and the source code of our computational approach are available at Github: https://github.com/JunaedYounusKhan51/COVID-19Base.

Limitations

Understandably, our findings through the knowledgebase prepared comes with some errors due to the inherent limitations of the methods and approaches adopted. This is why the identified inferences/associations are made available to the users for review to facilitate a feedback mechanism (from the esteemed users) in the COVID-19Base.

Conclusions

We have proposed a dictionary-based automated computational workflow to find the associations of six different thematic areas related to COVID-19/SARS-CoV-2 and other coronavirus-related diseases in humans and prepared a knowledgebase and made it available as a tool for the scientific community. We believe this knowledgebase would help the research community to explore the existing drugs and biomedical entities for coronavirus related diseases, and the lessons learned from the precedents of outbreak will allow us to find an effective treatment for COVID-19.

Acknowledgements

None.

Conflicts of Interest

None

Abbreviations

ACE2: Angiotensin-converting enzyme 2
CORD-19: COVID-19 Open Research Dataset
COVID-19: Coronavirus disease 2019
CRP: C-reactive protein
DNN: Deep Neural Network
DO: Disease Ontology
FDA: Food and Drugs Administration
HCV: Hepatitis C virus
HGNC: HUGO Gene Nomenclature Committee
ncRNAs: Non-coding RNAs

NIAID: National Institute of Allergy and Infectious Diseases
PDB: Protein Data Bank
POS: Part-of-Speech
ReLU: Rectified Linear Unit
SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2
tf-idf: Term frequency-inverse document frequency

Multimedia Appendix

The knowledgebase materials and the source code of our computational approach are available at Github: https://github.com/JunaedYounusKhan51/COVID-19Base.

References

1. Fanelli, D. and F. Piazza, *Analysis and forecast of COVID-19 spreading in China, Italy and France.* Chaos, Solitons & Fractals, 2020. **134**: p. 109761.
2. Chinazzi, M., et al., *The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak.* Science, 2020. **368**(6489): p. 395-400.
3. Heymann, D.L. and N. Shindo, *COVID-19: what is next for public health?* The Lancet, 2020. **395**(10224): p. 542-545.
4. de Wit, E., et al., *Middle East respiratory syndrome coronavirus (MERS-CoV) causes transient lower respiratory tract infection in rhesus macaques.* Proceedings of the National Academy of Sciences, 2013. **110**(41): p. 16598-16603.
5. Alam, I., et al., *Functional pangenome analysis suggests inhibition of the protein E as a readily available therapy for COVID-2019.* bioRxiv, 2020.
6. Muralidharan, N., et al., *Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 Protease against COVID-19.* Journal of Biomolecular Structure and Dynamics, 2020(just-accepted): p. 1-7.
7. Stebbing, J., et al., *COVID-19: combining antiviral and anti-inflammatory treatments.* The Lancet Infectious Diseases, 2020. **20**(4): p. 400-402.
8. Kandeel, M. and M. Al-Nazawi, *Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease.* Life sciences, 2020: p. 117627.
9. Lavecchia, A. and C. Di Giovanni, *Virtual screening strategies in drug discovery: a critical review.* Current medicinal chemistry, 2013. **20**(23): p. 2839-2860.
10. Li, Y., et al., *ViRBase: a resource for virus-host ncRNA-associated interactions.* Nucleic Acids Res, 2015. **43**(Database issue): p. D578-82.
11. Tang, D., et al., *VISDB: a manually curated database of viral integration sites in the human genome.* Nucleic acids research, 2020. **48**(D1): p. D633-D641.
12. Zhang, Y., et al., *Hepatitis C Virus Database and Bioinformatics Analysis Tools in the Virus Pathogen Resource (ViPR)*, in *Hepatitis C Virus Protocols*. 2019, Springer. p. 47-69.
13. Pickett, B.E., et al., *Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community.* Viruses, 2012. **4**(11): p. 3209-3226.
14. Lu Wang, L., et al., *CORD-19: The Covid-19 Open Research Dataset.* ArXiv, 2020.
15. Yates, B., et al., *Genenames. org: the HGNC and VGNC resources in 2017.* Nucleic acids research, 2016: p. gkw1033.

16. Goodsell, D.S., et al., *RCSB Protein Data Bank: Enabling biomedical research and drug discovery.* Protein Sci, 2020. **29**(1): p. 52-65.
17. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic acids research, 2007. **36**(suppl_1): p. D154-D158.
18. Schriml, L.M., et al., *Human Disease Ontology 2018 update: classification, content and workflow expansion.* Nucleic Acids Res, 2019. **47**(D1): p. D955-d962.
19. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
20. Kuhn, M., et al., *The SIDER database of drugs and side effects.* Nucleic acids research, 2016. **44**(D1): p. D1075-D1079.
21. Aho, A.V. and M.J. Corasick, *Efficient string matching: an aid to bibliographic search.* Communications of the ACM, 1975. **18**(6): p. 333-340.
22. Wang, P., et al., *Large-scale extraction of drug–disease pairs from the medical literature.* Journal of the Association for Information Science and Technology, 2017. **68**(11): p. 2649-2661.
23. Zimek, A., E. Schubert, and H.P. Kriegel, *A survey on unsupervised outlier detection in high-dimensional numerical data.* Statistical Analysis and Data Mining: The ASA Data Science Journal, 2012. **5**(5): p. 363-387.
24. Lloyd, S., *Least squares quantization in PCM.* IEEE transactions on information theory, 1982. **28**(2): p. 129-137.
25. Han, L. *Using a dynamic K-means algorithm to detect anomaly activities*. in *2011 Seventh International Conference on Computational Intelligence and Security*. 2011. IEEE.
26. Lima, M.F., et al. *Anomaly detection using baseline and k-means clustering*. in *SoftCOM 2010, 18th International Conference on Software, Telecommunications and Computer Networks*. 2010. IEEE.
27. Lu, W. and I. Traoré, *Unsupervised anomaly detection using an evolutionary extension of k-means algorithm.* International Journal of Information and Computer Security, 2008. **2**(2): p. 107-139.
28. Syarif, I., A. Prugel-Bennett, and G. Wills. *Unsupervised clustering approach for network anomaly detection*. in *International conference on networked digital technologies*. 2012. Springer.
29. Yasami, Y. and S.P. Mozaffari, *A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods.* The Journal of Supercomputing, 2010. **53**(1): p. 231-245.
30. Le, Q. and T. Mikolov. *Distributed representations of sentences and documents*. in *International conference on machine learning*. 2014.
31. Loria, S., et al., *Textblob: simplified text processing.* Secondary TextBlob: simplified text processing, 2014. **3**.
32. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.
33. Jones, K.S., *A statistical interpretation of term specificity and its application in retrieval.* Journal of documentation, 1972.
34. Luhn, H.P., *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of research and development, 1957. **1**(4): p. 309-317.
35. Pasupa, K. and W. Sunhem. *A comparison between shallow and deep architecture classifiers on small dataset*. in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*. 2016.
36. Feng, S., H. Zhou, and H. Dong, *Using deep neural network with small dataset to predict material defects.* Materials & Design, 2019. **162**: p. 300-310.
37. Hahnloser, R.H., et al., *Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit.* Nature, 2000. **405**(6789): p. 947-951.

38. Karlik, B. and A.V. Olgac, *Performance analysis of various activation functions in generalized MLP architectures of neural networks.* International Journal of Artificial Intelligence and Expert Systems, 2011. **1**(4): p. 111-122.
39. Menon, A., et al., *Characterization of a class of sigmoid functions with applications to neural networks.* Neural Networks, 1996. **9**(5): p. 819-835.
40. Glorot, X. and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*. in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010.
41. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
42. Singhal, A., *Modern information retrieval: A brief overview.* IEEE Data Eng. Bull., 2001. **24**(4): p. 35-43.
43. Piñero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic acids research, 2016: p. gkw943.
44. Wang, J.T., et al., *Clinical manifestations, laboratory findings, and treatment outcomes of SARS patients.* Emerg Infect Dis, 2004. **10**(5): p. 818-24.
45. Chen, C., et al., *[Analysis of myocardial injury in patients with COVID-19 and association between concomitant cardiovascular diseases and severity of COVID-19].* Zhonghua Xin Xue Guan Bing Za Zhi, 2020. **48**(0): p. E008.
46. Wang, G., et al., *C-Reactive Protein Level May Predict the Risk of COVID-19 Aggravation.* Open Forum Infectious Diseases, 2020. **7**(5).
47. Akgun, E., et al., *ALTERED MOLECULAR PATHWAYS OBSERVED IN NASO-OROPHARYNGEAL SAMPLES OF SARS-CoV-2 PATIENTS.* medRxiv, 2020.
48. Potì, F., et al., *Treatments for COVID-19: emerging drugs against the coronavirus.* Acta Biomed, 2020. **91**(2): p. 118-136.
49. Walls, A.C., et al., *Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein.* Cell, 2020.
50. Hoffmann, M., et al., *SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor.* Cell, 2020.
51. Zang, R., et al., *TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes.* Sci Immunol, 2020. **5**(47).
52. Wicik, Z., et al., *ACE2 interaction networks in COVID-19: a physiological framework for prediction of outcome in patients with cardiovascular risk factors.* BioRxiv, 2020.
53. He, R., et al., *The clinical course and its correlated immune status in COVID-19 pneumonia.* Journal of Clinical Virology, 2020: p. 104361.
54. D'Amico, F., et al., *Diarrhea during COVID-19 infection: pathogenesis, epidemiology, prevention and management.* Clinical Gastroenterology and Hepatology, 2020.
55. Horby, P., et al., *Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report.* medRxiv, 2020: p. 2020.06.22.20137273.
56. Chaccour, C., et al., *Ivermectin and COVID-19: Keeping Rigor in Times of Urgency*, in *Am J Trop Med Hyg*. 2020. p. 1156-1157.
57. Caly, L., et al., *The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro.* Antiviral Res, 2020. **178**: p. 104787.
58. Sujan, M.A. *Use of Ivermectin: Hope held out, caution called for*. 2020 [cited 2020 26 June]; Available from: https://www.thedailystar.net/frontpage/news/use-ivermectin-hope-held-out-caution-called-1914041.
59. Al-Tawfiq, J.A., A.H. Al-Homoud, and Z.A. Memish, *Remdesivir as a possible therapeutic option for the COVID-19.* Travel Med Infect Dis, 2020. **34**: p. 101615.
60. Tchesnokov, E.P., et al., *Mechanism of Inhibition of Ebola Virus RNA-Dependent RNA Polymerase by Remdesivir.* Viruses, 2019. **11**(4).
61. Warren, T.K., et al., *Therapeutic efficacy of the small molecule GS-5734 against*

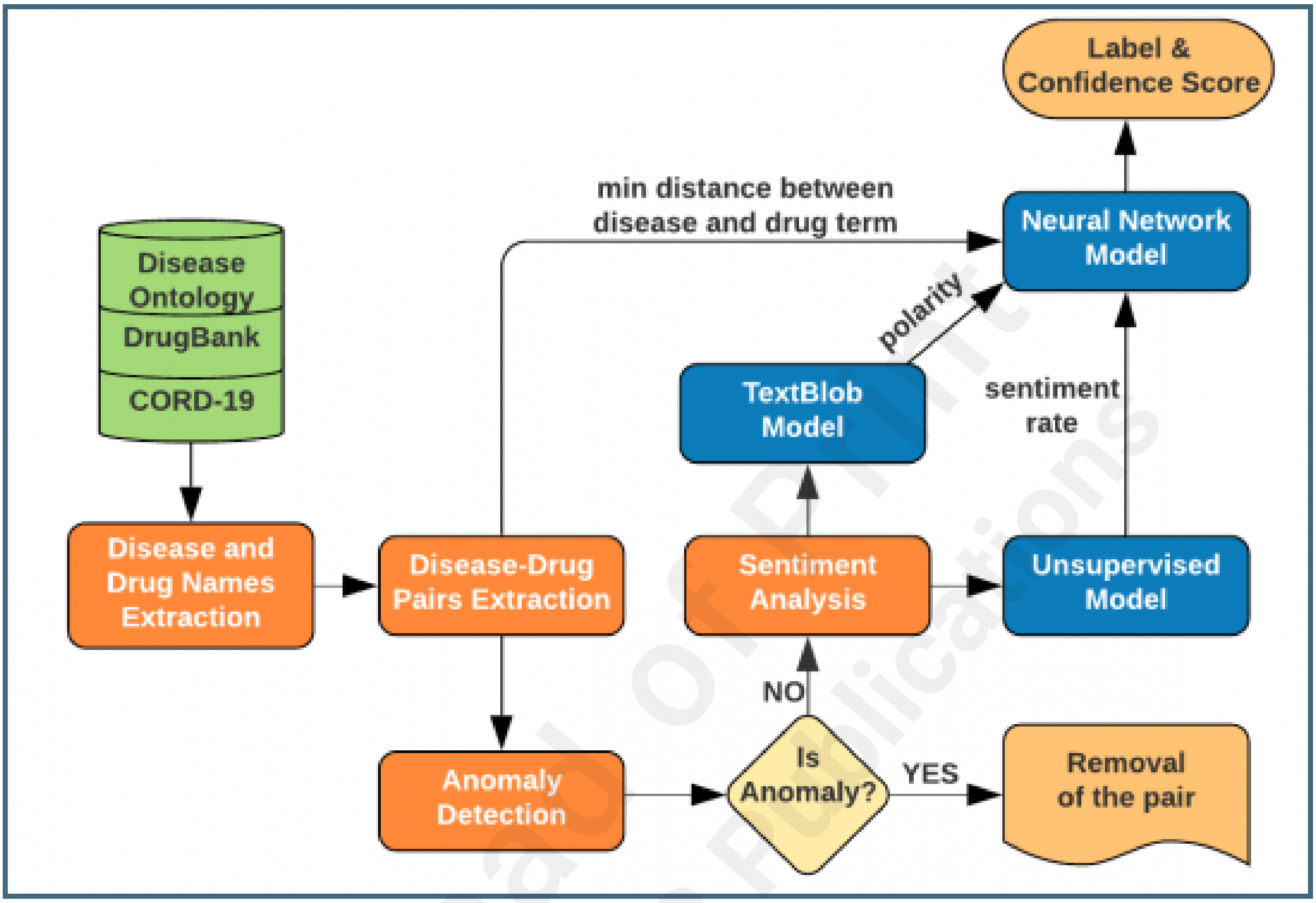
Ebola virus in rhesus monkeys. Nature, 2016. **531**(7594): p. 381-5.

62. *NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19*. 2020 [cited 2020 26 June]; Available from: https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19.
63. *Clinical trials related to COVID-19*. 2020 [cited 2020 26 June]; Available from: https://clinicaltrials.gov/ct2/results?cond=COVID-19.
64. Hagar, M., et al., *Investigation of Some Antiviral N-Heterocycles as COVID 19 Drug: Molecular Docking and DFT Calculations.* Int J Mol Sci, 2020. **21**(11).
65. Grein, J., et al., *Compassionate Use of Remdesivir for Patients with Severe Covid-19.* N Engl J Med, 2020. **382**(24): p. 2327-2336.
66. Wang, Y., et al., *Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial.* Lancet, 2020. **395**(10236): p. 1569-1578.
67. *FDA cautions against use of hydroxychloroquine or chloroquine for COVID-19 outside of the hospital setting or a clinical trial due to risk of heart rhythm problems*. 2020 [cited 2020 12 May]; Available from: https://www.fda.gov/drugs/drug-safety-and-availability/fda-cautions-against-use-hydroxychloroquine-or-chloroquine-covid-19-outside-hospital-setting-or.
68. Cannon, C.P., et al., *Intensive versus moderate lipid lowering with statins after acute coronary syndromes.* N Engl J Med, 2004. **350**(15): p. 1495-504.
69. Dashti-Khavidaki, S. and H. Khalili, *Considerations for Statin Therapy in Patients with COVID-19.* Pharmacotherapy, 2020. **40**(5): p. 484-486.
70. Castiglione, V., et al., *Statin therapy in COVID-19 infection.* Eur Heart J Cardiovasc Pharmacother, 2020.
71. *Coronavirus Response - Active Support for Hospitalised Covid-19 Patients (CRASH-19)*. 2020 [cited 2020 26 June]; Available from: https://clinicaltrials.gov/ct2/show/NCT04343001.
72. *Atorvastatin as Adjunctive Therapy in COVID-19 (STATCO19)*. 2020 [cited 2020 26 June]; Available from: https://clinicaltrials.gov/ct2/show/NCT04380402.
73. Deliwala, S., et al., *Encephalopathy as the Sentinel Sign of a Cortical Stroke in a Patient Infected With Coronavirus Disease-19 (COVID-19).* Cureus, 2020. **12**(5): p. e8121.
74. Alexander, S.P.H., et al., *A rational roadmap for SARS-CoV-2/COVID-19 pharmacotherapeutic research and development. IUPHAR Review 29.* Br J Pharmacol, 2020.
75. Ottosen, S., et al., *In vitro antiviral activity and preclinical and clinical resistance profile of miravirsen, a novel anti-hepatitis C virus therapeutic targeting the human factor miR-122.* Antimicrobial agents and chemotherapy, 2015. **59**(1): p. 599-608.
76. Shi, Y., et al., *COVID-19 infection: the perspectives on immune responses*, in *Cell Death Differ*. 2020. p. 1451-1454.
77. Amanat, F. and F. Krammer, *SARS-CoV-2 Vaccines: Status Report.* Immunity, 2020. **52**(4): p. 583-589.
78. Rosa, S.G.V. and W.C. Santos, *Clinical trials on drug repositioning for COVID-19 treatment.* Rev Panam Salud Publica, 2020. **44**: p. e40.
79. Remmelts, H.H., et al., *Biomarkers define the clinical response to dexamethasone in community-acquired pneumonia.* J Infect, 2012. **65**(1): p. 25-31.
80. da Costa, D.E., et al., *Steroids in full term infants with respiratory failure and pulmonary hypertension due to meconium aspiration syndrome.* Eur J Pediatr, 2001. **160**(3): p. 150-3.

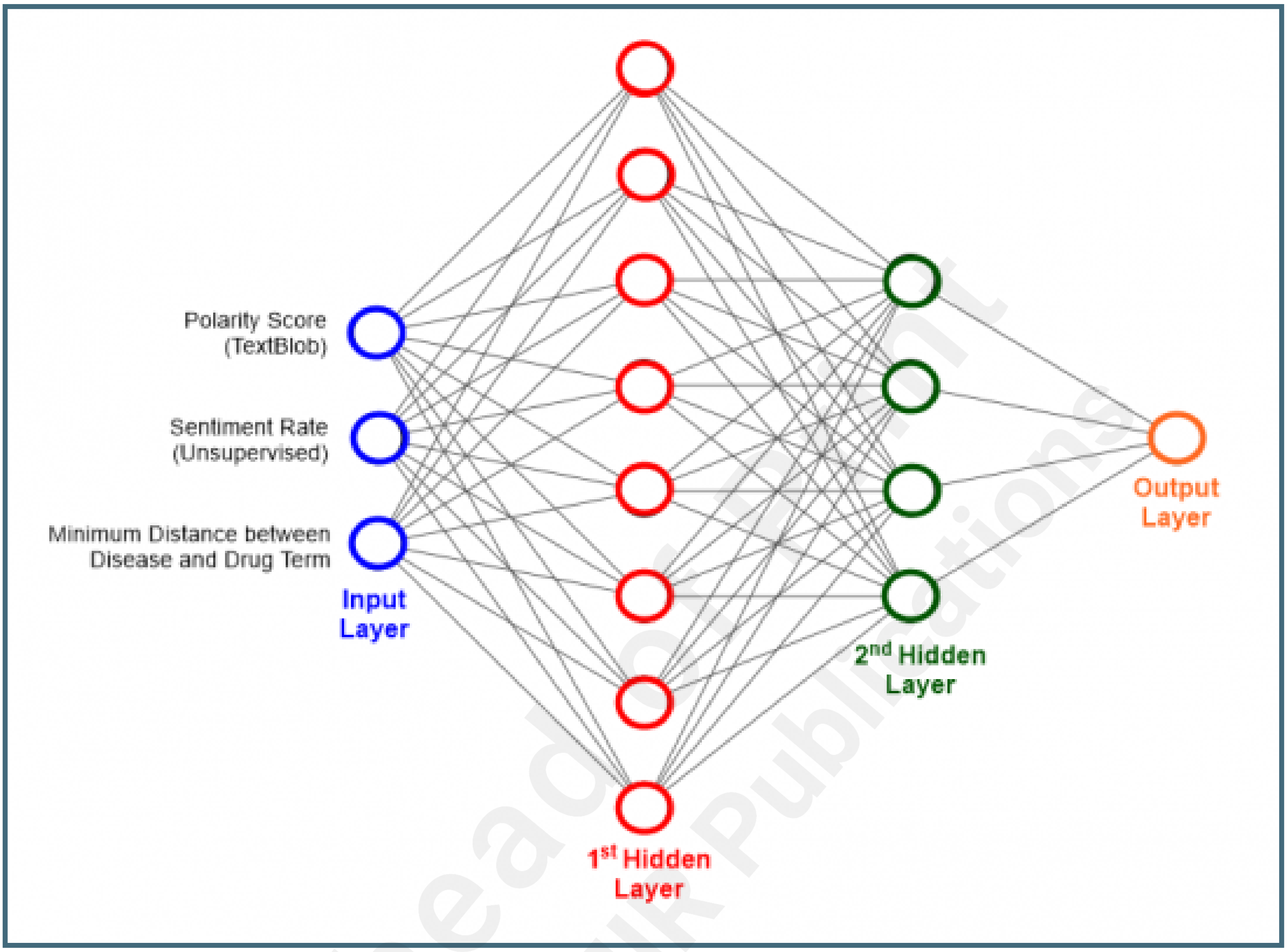
Supplementary Files

Figures

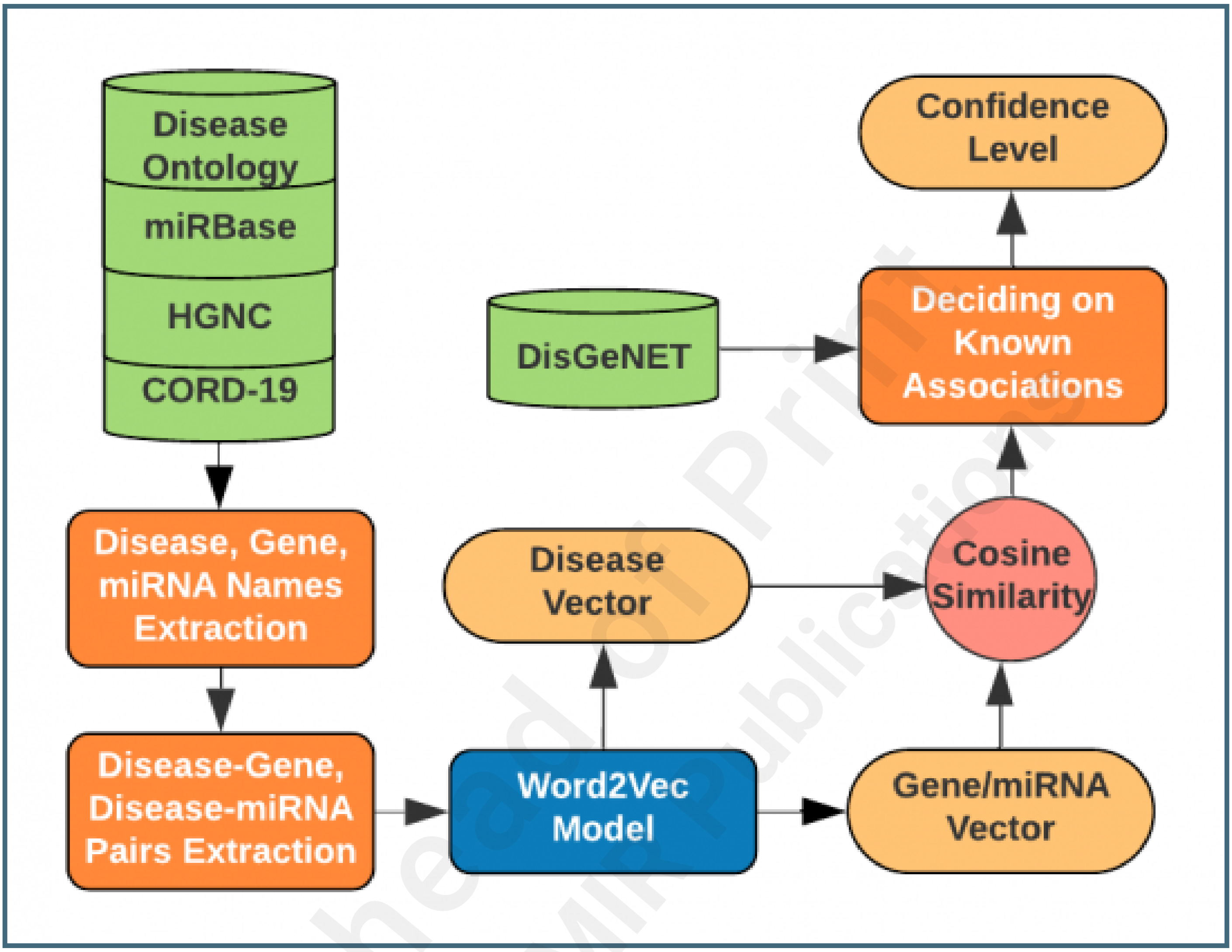
Flowchart of extracting disease-drug interactions and predicting the effectiveness of drugs against diseases with confidence scores.



Schematic diagram of the deep neural network used to predict the effectiveness of drugs against diseases.



Flowchart of extracting disease-gene, disease-miRNA associations and determining their confidence levels.


Disease Ontology
miRBase
HGNC
CORD-19
DisGeNET
Confidence Level
Deciding on Known Associations
Disease, Gene, miRNA Names Extraction
Disease Vector
Cosine Similarity
Disease-Gene, Disease-miRNA Pairs Extraction
Word2Vec Model
Gene/miRNA Vector