Journal of Taibah University Medical Sciences (2023) 18(4), 787–801

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com

Original Article

Cluster-based text mining for extracting drug candidates for the prevention of COVID-19 from the biomedical literature

Ahmad Afif Supianto, Ph.D [a,‡], Rizky Nurdiansyah, M.Si [b,‡], Chia-Wei Weng, Ph.D [c,‡], Vicky Zilvan, M.T. [a,#], Raden Sandra Yuwana, M.T. [a,#], Andria Arisal, MEDC [a,#], Hilman Ferdinandus Pardede, Ph.D [a,#], Min-Min Lee, Ph.D [d], Chien-Hung Huang, Ph.D [e] and Ka-Lok Ng, Ph.D [f,g,h,*]

[a] *Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia*
[b] *Department of Bioinformatics, Indonesia International Institute for Life Sciences, Indonesia*
[c] *Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan*
[d] *Department of Food Nutrition and Health Biotechnology, Asia University, Taiwan*
[e] *Department of Computer Science and Information Engineering, National Formosa University, Taiwan*
[f] *Department of Bioinformatics and Medical Engineering, Asia University, Taiwan*
[g] *Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan*
[h] *Center for Artificial Intelligence and Precision Medicine Research, Asia University, Taiwan*

Received 11 June 2022; revised 14 October 2022; accepted 12 December 2022; Available online 4 January 2023

الملخص

أهداف البحث: جعلت الأزمة الصحية كوفيد-19 التي بدأت في نهاية عام 2019 الباحثين من جميع أنحاء العالم يتسابقون بسرعة لإيجاد حلول فعالة حتى الآن. كثرت الأبحاث ذات الصلة وكان من المحتم أن تكون هناك حاجة إلى نهج آلي للعثور على معلومات مفيدة ، وبالتحديد التنقيب عن النص ، للتغلب على كوفيد-19، لا سيما فيما يتعلق باكتشاف مرشح العلاج. بينما تحاول طرق التنقيب عن النص للعثور على الأدوية المرشحة في الغالب استخراج ارتباطات حيوية من "بابميد"، إلا أن القليل جدا منها يستخدم أسلوب التجميع. الغرض من البحث هو إثبات فعالية نهجنا في تحديد الأدوية للوقاية من كوفيد-19 من خلال مراجعة الأبحاث وتحليل الكتلة وحسابات إرساء الأدوية وبيانات التجارب السريرية.

طريقة البحث: تم إجراء هذا البحث في أربع مراحل رئيسية. أولا، تم تنفيذ مرحلة التنقيب عن النص من خلال إشراك "بايوبيرت" للحصول على تمثيل متجه لكل كلمة في الجملة من النصوص. كانت المرحلة التالية هي إنشاء روابط دوائية للأمراض يتم الحصول عليها من المراسلات بين المرض والعقار. بعد ذلك ، جمعت مرحلة التجميع القواعد من خلال تشابه الأمراض من خلال استخدام "تي إف-آي دي إف" كميزات لها. أخيرا، تتم معالجة مرحلة استخراج مرشح الدواء من خلال الاستفادة من قواعد بيانات "بابكيم" و "بنك الدواء". كما استخدمنا حزمة إرساء الأدوية "أوتودوك فينا" في برنامج "بي واي آر إكس" للتحقق من النتائج.

النتائج: أظهر التحليل المقارن الذي تم إجراؤه أن النسبة المئوية للنتائج المستخدمة في التعدين مع العنقودية تفوقت على التعدين دون التجميع في جميع البيئات التجريبية. بالإضافة إلى ذلك ، اقترحنا أن أفضل ثلاثة أدوية / مواد كيميائية نباتية من خلال تحليل الالتحام بالعقاقير قد تكون فعالة في الوقاية من كوفيد-19.

الاستنتاجات: تعد الطريقة المقترحة لتعدين النص باستخدام طريقة التجميع واعدة للغاية في اكتشاف الوقاية من الأدوية المرشحة لكوفيد-19 من خلال الأدبيات الطبية الحيوية.

الكلمات المفتاحية: فيروس كورونا؛ كوفيد-19؛ سارس-كوف-2؛ تحليل النصوص؛ إرساء الدواء؛ المواد الكيميائية النباتية

Abstract

Objective: The coronavirus disease 2019 (COVID-19) health crisis that began at the end of 2019 made researchers around the world quickly race to find effective solutions. Related literature exploded and it was inevitable that an automated approach was needed to find useful information, namely text mining, to overcome COVID-19,

* Corresponding address: Department of Bioinformatics and Medical Engineering, No. 500, LiuFeng Rd., WuFeng Dist., Taichung City, 41354, Taiwan.
E-mail: ppiddi@gmail.com (K.-L. Ng)

Peer review under responsibility of Taibah University.

Production and hosting by Elsevier

‡ First authors, equal contribution.
# Second authors, equal contribution.

 https://doi.org/10.1016/j.jtumed.2022.12.015

especially in terms of drug candidate discovery. While text mining methods for finding drug candidates mostly try to extract bioentity associations from PubMed, very few of them mine with a clustering approach. The purpose of this study was to demonstrate the effectiveness of our approach to identify drugs for the prevention of COVID-19 through literature review, cluster analysis, drug docking calculations, and clinical trial data.

Methods: This research was conducted in four main stages. First, the text mining stage was carried out by involving Bidirectional Encoder Representations from Transformers for Biomedical to obtain vector representation of each word in the sentence from texts. The next stage generated the disease-drug associations, which were obtained from the correlation between disease and drug. Next, the clustering stage grouped the rules through the similarity of diseases by utilizing Term Frequency-Inverse Document Frequency as its feature. Finally, the drug candidate extraction stage was processed through leveraging PubChem and DrugBank databases. We further used the drug docking package AUTODOCK VINA in PyRx software to verify the results.

Results: Comparative analyses showed that the percentage of findings using mining with clustering outperformed mining without clustering in all experimental settings. In addition, we suggest that the top three drugs/phytochemicals by drug docking analysis may be effective in preventing COVID-19.

Conclusions: The proposed method for text mining utilizing the clustering method is quite promising in the discovery of drug candidates for the prevention of COVID-19 through the biomedical literature.

Keywords: Coronavirus; COVID-19; Drug docking; Phytochemicals; SARS-CoV-2; Text mining

Introduction

Ever since the coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus struck in late 2019, researchers worldwide have studied the disease and raced towards finding a cure and developing effective treatments. This has caused an explosion of biomedical literature regarding the disease, its virus, and potential drug candidates. Several traditional literature review surveys have been proposed on the COVID-19 pandemic 19[1–4] focusing on several key issues including symptoms, the coincidence of COVID-19 and other diseases, and various treatments. However, these studies have limitations, particularly in that the literature reviews are often topic-specific, limited to a few diseases or chemicals, as well as time-consuming and labor-intensive. Therefore, an automated approach is needed to find useful information to overcome COVID-19, especially in terms of drug candidate discovery.

One literature database known as PubMed, developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), one of the institutes of the National Institutes of Health (NIH), has published about 100,000 articles discussing COVID-19, and as of April 2021, with a growth rate of dozens of new articles every day. This huge amount of text data once again shows that there is a need to explore a collection of studies with automated text mining approaches to find potential drug candidates for the COVID-19 pandemic.

Several text mining studies have utilized computational methods to find drug candidates. Kuusisto et al. proposed a word embedding approach built on biomedical literature published through early 2019.[5] To obtain a list of drugs, the study downloaded the United States Food and Drug Administration- (FDA) approved drug databases, extracted drug names, and processed them for use in word embedding. Muramatsu and Tanokura applied the relationships between pairs of terms in PubMed abstracts and used the co-occurrences.[6] They measured the distance between two Kyoto Encyclopedia of Genes and Genomes (KEGG) codes determined by the number of articles (PubMed ID) where terms from the KEGG pair appeared together.

Specifically, drug discovery is carried out by evaluating the disease-drug associations. Chen et al. applied a combination of Natural Language Processing (NLP) and statistical techniques for the automated acquisition of disease-drug associations in Medline articles.[7] Disease and drug entities were identified using the NLP systems in addition to Medical Subject Heading annotations for the Medline articles. Co-occurrence statistics were then applied to compute and evaluate the strength of association between each disease and relevant drugs. However, drug checking by involving compound entities and integrating with chemical database repositories was not carried out. This study used PubChem to validate the findings of drug candidates to ensure that these compounds have a standard representation of compounds.

In another body of research with a different approach, Khan et al. extracted disease-drug interactions from the CORD-19 literature and classified them into positive or negative labels.[8] The positive label means the drug is potentially effective against COVID-19, and the negative label means the opposite. In a similar way, Wang et al. developed a computational drug repositioning approach to discover potential drug-disease associations. They applied an ensemble strategy to predict the association of drug-disease pairs based on improved drug-disease association information and the constructed similarity network.[9] Instead of classifying and predicting the disease-drug associations, we grouped similar disease terms into clusters, and then using a frequency analysis approach, we registered candidate drugs to search for their presence in the COVID-19 DrugBank database.

This research provides a novel data-driven framework to analyze COVID-19-related papers to find COVID-19 drug candidates. The aims of this study were to (i) show that the clustering method has a positive impact on the discovery of

COVID-19 drug candidates; and (ii) use drug docking calculations and clinical trial data to demonstrate the effectiveness of our approach.

Materials and Methods

Dataset

This study used three open datasets that are available and can be accessed online. First, PubMed was used as a data source to obtain articles. PubMed (https://pubmed.ncbi.nlm.nih.gov) is a database designed to provide access to citations (with abstracts) from valuable sources of information for scientific research in biomedical literature. PubMed comprises 32 million citations developed by the NCBI at the National Library of Medicine, one of the institutes of the NIH. PubMed was chosen because of the rapid growth of biomedical literature—about two articles are added every minute on average. PubMed has been used by several studies related to text mining such as a text mining tool that enables users to perform classification of scientific literature by text mining-based classification of PubMed article abstracts,[10] predict gene−disease associations from the texts of documents in the PubMed database,[11] text mining tools for extracting information from PubMed abstracts of scientific papers in the food microbiology domain,[12] and text mining framework for interactive analysis and visualization of similarities among biomedical entities.[13]

The second dataset, PubChem (https://pubchem.ncbi.nlm.nih.gov), was used as a source to validate drug compounds obtained from the drug candidates extraction stage. PubChem is a popular public repository for chemical substance information resources and their biological activities that serves the scientific community as well as the general public by the NIH. Many studies have used PubChem, including those that are used to train machine learning/deep learning algorithms, which are used to predict the drug targets,[14] in order to uncover and summarize important relationships among chemicals, genes, proteins, and diseases by analyzing co-occurrences of terms in biomedical literature abstracts,[15] and to generate a comprehensive blood exposome database of endogenous and exogenous chemicals associated with the mammalian circulatory system through text mining and database fusion.[16]

The third dataset is DrugBank (https://go.drugbank.com/covid-19), which is used as a source of validation as to whether the drug candidate is suitable to treat COVID-19. The DrugBank database has been used to explore drug associations in building a text mining tool for knowledge discovery,[17] drug name recognition that deals with different types of drug−drug interactions texts and language styles,[18] and to construct a protein vector to assess the effectiveness of similarity-based drug−drug interaction prediction.[19]

Method

We extracted scientific articles from PubMed using Bidirectional Encoder Representations from Transformers for Biomedical (BioBERT) Text Mining, established disease-drug associations, grouped similar disease-drug associations using agglomerative hierarchical clustering, checked the availability of drug candidates on PubChem through compound validation, and found a list of drug candidates on the DrugBank COVID-19. The process flow of the proposed method is shown in Figure 1. The method is divided into four main stages: the text mining stage, generating rule stage, clustering stage, and extracting potential drug stage. The text mining stage involved BioBERT as the text mining tool and PubMed as the data source. The generating rules stage proposed ideas by building rules obtained from the correspondence between the disease and drug found in the abstract of the articles. The clustering stage grouped rules that have similarities to text disease by using Term Frequency-Inverse Document Frequency (TF-IDF) as the feature. We proposed agglomerative hierarchical clustering (AHC) with the output in the form of a list of rules as cluster members. This list of rules is used as input for the last stage, namely the extracting potential drug stage. This extraction stage utilized PubChem as a validation source for drug compounds and is continued by utilizing DrugBank as a validation source that the drug compounds have the potential to treat COVID-19. Two evaluations were then applied to determine the potential COVID-19 drug candidates, which are percentage of drugs found in DrugBank.

Text mining stage

The initial stage of our proposed method is the determination of search keywords related to COVID-19. The text mining technique used in this study starts from two sets of keywords: keywords for human diseases caused by viruses (SARS, coronavirus, human immunodeficiency virus [HIV], Middle East respiratory syndrome, and Ebola) along with Chinese medicinal herb compound names. Chinese herbal medicine has achieved significant clinical efficacy in the interventional treatment of human diseases. Yang et al. analyzed the etiology of COVID-19 and the efficacy of clinical Chinese medicine active ingredients.[20] During the course of COVID-19 treatment, traditional Chinese medicine has antiviral antivirus, anti-inflammatory, and immunoregulation effects on treating COVID-19[21]; hence, this motivated us to establish search keywords using Chinese herbal medicine.

The next step was to search for articles on the PubMed website based on those keywords, and the result was a list of PubMed article IDs, which was then used as input to obtain information on disease and potential drugs using BioBERT text mining.[22] BioBERT has proven to be a successful tool for biomedical linkage extraction. It has been used to optimize biomedical relationship extraction in identifying functional links between proteins,[23] and to obtain vector representation of each word in the sentence in extracting drug−drug interactions from the texts.[24] The output of this stage is a list of articles along with seven categories of information including gene/protein, disease, drug/chemicals, species, mutations, microRNAs and pathways.

Generating rules stage

The search for drug candidates is based on the idea of building association rules. Rules are built on the basis of the

correspondence relationship between "disease" and "drug." Before the rule is built, articles that do not contain disease or drugs will be removed from the input list from the generating rules process. Through correspondence between them, a pair of disease and drug candidates will be obtained. If there is a rule $X \rightarrow Y$, it means that if there is a disease X, then the drug candidate is Y. In an article, it is possible that there is more than one disease or one drug. We then combined the entire list of diseases and drugs in the article, so that a unique rule will be formed in an article. This process will be carried out on all articles obtained in the previous stage.

We believe that a rule does not only occur in an article. Therefore, we adopted the term 'support' in the association rule mining concept to indicate the frequency of these rules. One rule in an article represents one support. If there is the same rule in another article, then that rule gets one additional support. In other words, if a rule is found in five articles, it means that the support for that rule is five. We call this approach "frequency analysis." Frequency analysis plays an important role in text mining. We utilized frequency analysis to understand the significance of the rule. This analysis showed how many articles contained the disease and its drug treatment.

The last step of this stage is filtering. Filtering needs to be done because not all the rules that are built are related to COVID-19 even though the articles searched use the COVID-19 keyword. BioBERT is a text mining tool for disease and drug identification in general, not specifically for COVID-19 disease and drug detection. Therefore, in the final step we applied filtering criteria to the rules with the same keywords as in the article search.

Clustering stage

The output of the rules-generating process is a list of rules that specify the disease-drug association. We found some disease words that have similar meanings but are written in different ways. For instance, in writing "sars cov-2," we found that in some articles it was written as "sars- cov-2," "sars-cov-2," "sars-cov- 2," "sars-cov 2," and "sars-cov-2-." To deal with this condition, we proposed implementing text clustering to group similar texts, which means similar diseases. This approach was carried out with the argument that if these similar terms occur in many articles with different rules, then the rule with the similar text will have low support. Rules with low support will be considered rules with low interest. On the other hand, it is different when we collect similar rules into one and add up the support; then the support for that rule will be higher and have the potential to become an interesting rule. This interesting rule will potentially become a candidate for the discovery of a COVID-19 drug. For instance, the rules with the drug candidate of azithromycin are follows:

- sars-cov-2 → azithromycin: supported by 6 articles
- sars-cov-2- → azithromycin: supported by 1 article
- sars-cov → azithromycin: supported by 1 article
- sars-cov-2 infection -> azithromycin: supported by 2 articles
- sars-cov-2 virus -> azithromycin: supported by 1 article

Following those rules of grouping, they would be supported by 11 articles, instead of only 1, 2, or 6 articles respectively. We adopted AHC to group the rules, which generally follow the five main steps listed below:

1. Calculating the distance matrix for the initial clusters using cosine similarity.
2. Searching for the minimum distance in the matrix.
3. Combining the two clusters with the minimum distance.
4. The distance matrix can be updated by calculating the distances between the new cluster with the other clusters.
5. Repeating the previous three steps if more than one cluster remains.

Extracting potential drugs stage

The disease-drug relationship that was selected in the previous stage was calculated for its frequency, and the ones with higher frequency were selected. We introduced the term 'minimum support' as a parameter that indicates the minimum number of frequencies of the rules that will be selected and are considered as interesting rules. The higher the minimum support value, the less the number of rules that will be obtained. This will make the rules interesting because of many articles related to it. However, the few rules obtained resulted in fewer drug candidates identified.

After obtaining the rules that meet the minimum support criteria, we checked the availability of drug candidates on PubChem. The results of this examination are properties of drug candidate compounds, in which we used the Simplified Molecular Input Line Entry System (SMILES), which was proposed in[25] and is widely used as a standard representation of compounds for chemical information processing.

Extracting the COVID-19 keywords from articles certainly does not guarantee that the drug is a candidate drugs for COVID-19. Therefore, validation of compounds to the COVID-19 dataset was carried out. Then the list of SMILES found on PubChem was searched on DrugBank COVID-19. The results found in DrugBank were the final findings as an outcome of the proposed method. We then calculated the percentage of these findings as an evaluation method for the success of our method.

Evaluation metrics

We used three types of evaluations: measuring (i) the quality of the dendrogram in the hierarchical clustering method, (ii) the quality of the cluster, and (iii) the percentage of drug findings. The measurement of dendrogram quality aimed to obtain the best performance setting of the linkage method in hierarchical clustering, and the Cophenetic correlation coefficient[26] was used. The second evaluation metric was a measure of cluster quality. This measurement was intended to obtain the best cutoff value in the dendrogram, which has an impact on the large number of clusters formed. We used the Davis-Bouldin index[27] to measure the quality of the cluster.

We used the ratio of the number of COVID-19 drug candidates found on DrugBank compared to the number of drug candidates found on PubChem as an indicator. Suppose that the number of drug candidates found in DrugBank and PubChem are X and Y, respectively, then the percentage of COVID-19 drug findings *(PF)* is defined by

$$PF = \frac{X}{Y} \times 100\% \quad (1)$$

Drug docking calculation with structures of SARS-CoV-2 proteins

To evaluate the drug candidates, a total of four structures of SARS-CoV-2 spike proteins were adopted for the docking calculation. The protein data bank (PDB) code of the receptor-binding domain (RBD) is 6VW1, (2.68 Å resolution, structure of SARS-CoV-2 chimeric receptor-binding domain complexed with its receptor human angiotensin-converting enzyme 2 [ACE2]).[28] Since the crystal structure was retrieved during the bond to its protein target (ACE2), it is suitable for docking analysis. Another protein we chose is the post-fusion core of 2019-nCoV S2 subunit (6LXT, 2.90 Å resolution). The crystal structure was made during inhibition with a fusion inhibitor, so it will be in the structure when binding to a certain target. The S2 subunit of 6LXT is able to facilitate the virus to penetrate the host cell membrane.[29] In addition, two SARS-CoV-2 spike ectodomain structures were also adopted in this docking analysis, including the open state (6VYB, 3.20 Å resolution) and closed state (6VXX, 2.80 Å resolution).[30] The docking procedure was based on a previous published study.[31] The water molecule and ligand molecule of the retrieved co-crystalized structure were removed using the PyMOL version 2.0.4 program to prepare for the docking analysis.

The chemical structures of the drug candidates were downloaded from the PubChem (SMILES, as saved in Surface Data File format). Each drug candidate was optimized in molecular geometry, torsional barriers, and intermolecular-interaction geometry using the MMFF94 partial forcefield in CHARMM. The structures of viral proteins were treated using the same forcefield during the docking calculation. BIOVIA Discovery Studio 2021 was used to apply the forcefield. We performed the drug docking calculation by using the AUTODOCK VINA package[32] in the PyRx software (https://pyrx.sourceforge.io/home) with the "maximize" setting in the grid selection. The best docking conformation for each drug candidate was determined based on the lowest binding energy. The binding energy for each drug candidate was adopted as the indicator to evaluate the binding affinity between the SARS-CoV-2 protein and drug candidate. Lower binding energy means the higher binding affinity between the SARS-CoV-2 protein and drug candidate, and vice versa. Post-docking analysis was determined and visualized using BIOVIA Discovery Studio 2021.

Experimental settings

We used the NLTK, Sklearn, SciPy, and PubChempy library with the Jupyter Notebook to implement our method. The code is written in Python 3.8.5. In our method, we used AHC, which has parameters that must be defined. We tested these parameters to obtain parameters that are able to provide optimal clustering results. For instance, we tested Single, Complete, Average, Weighted, Median, and Ward Linkages as the linkage measurement methods. A cophenetic coefficient was performed to compare the dendrogram generated from the linkage methods. Then, for the cutoff parameter, we use a ranged from 2.5 to 3.5 at a step of 0.1. Distances between data objects are given by the Euclidean distance.

We focused on rules with fairly high support, which means that many articles discussed the relationship between the disease and the COVID-19 drug candidates. The question is then whether the rules in the cluster with the highest support are sufficient to represent drug candidate discovery? Or do they still need rules in other clusters with lower support? We defined the question as the first research question (RQ1).

Regardless of what the answer is in RQ1, the result of the rules obtained will then be selected based on the specified minimum number of supports. Minimum support that is too low will result in many rules, but if minimum support is too high, it will result in just a few rules. The few rules will result in fewer drug candidates to be searched for. We suggested recommending more drug candidates. Since this is highly uncertain and the determination of the minimum support will have an impact on the discovery of drug candidates, we determined the parameter of minimum support values in the range from 3 to 12. A summary of the experimental settings is shown in Table 1. After the optimal parameters are obtained, referring to our argument by proposing the clustering approach in mining drug candidates for COVID-19, we tested our proposed method and asked the next research question, as to whether the findings of drug candidates with clustering are better than those without clustering (RQ2).

Results

Text mining and generating rules results

Based on the search results on the PubMed website with the specified keyword list, we obtained 26,327 articles. Articles were obtained from 1959 to January 2022. Using BioBERT, we obtained a list of diseases and drugs based on article IDs. Of the 26,327 articles obtained, 15,010 articles contained information on disease and drugs. The 15,010 articles were then used as the basis for generating rules and obtained 293,315 rules. Many established rules are not specifically related to COVID-19 such as: lung disease → *amino acid*, diabetes → *aldosterone*, influenza → *losartan*, and cardiac complications → *catecholamines*. We filtered it using the same keywords as in the first step of the PubMed search and found 4325 rules. Next, we listed diseases in the rules for clustering.

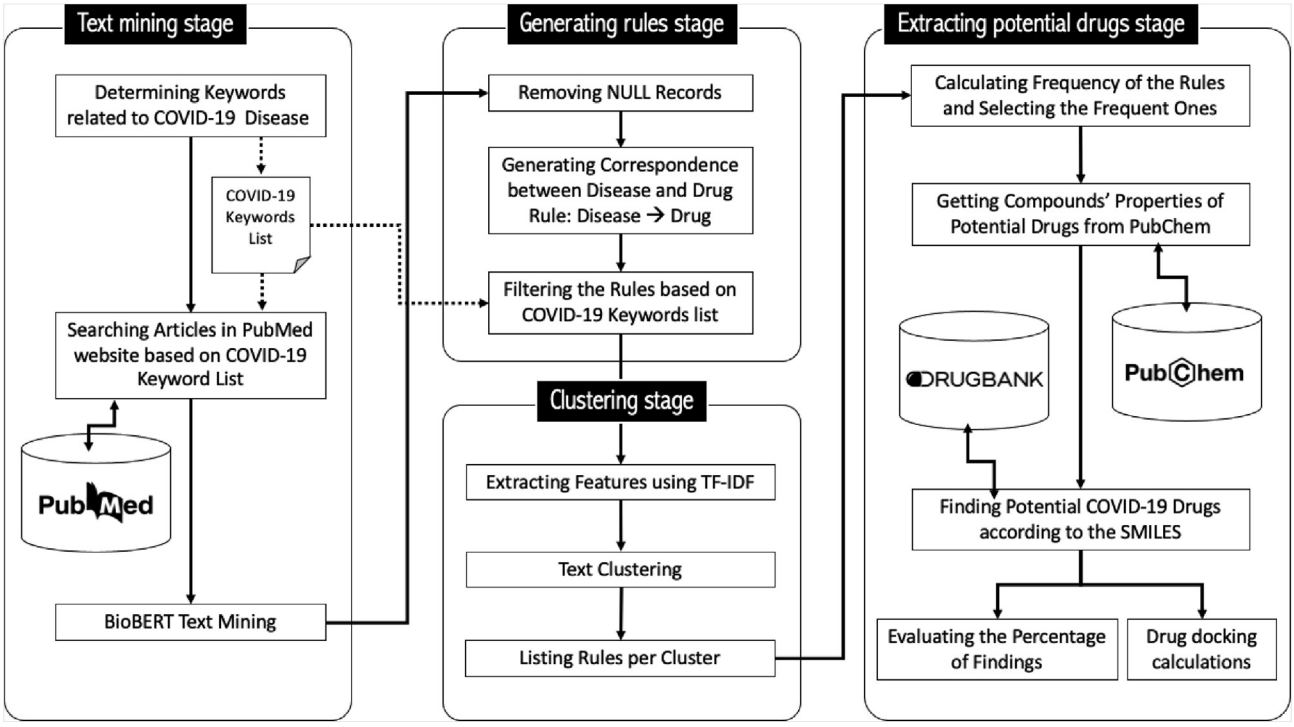

Figure 1: Process flow of the proposed method. Solid arrows: show process flow. Double solid arrows: indicate flow in−out system with open online dataset. Dash arrows: show system in−out flow with storage files.

Table 1: Experimental settings.

Method	Parameter	Experiments Value
AHC	linkage	Single, Complete, Average, Weighted, Median and Ward
	cutoff	2.5 to 3.5
	distance	Euclidean
Frequent Items	minsupp	3 to 12

The best performance setting of AHC

The results of the cophenetic correlation coefficient of the linkage methods are given in Table 2. The average linkage method yields the highest coefficient and outperforms other methods with a score of 0.7650. Next, we used average linkage to determine the best cutoff parameter. Based on the Davis-Bouldin score, it was found that the cutoff with a value of 2.6 yielded the smallest score of 1.0247. This meant that the cutoff parameter 2.6 was able to produce the best cluster formation. Details of the scores for each cutoff are listed in Table 3. With the selected cutoff value, we obtained 29 clusters. Further, the cluster analysis process is carried out using the best clustering method settings.

Table 2: The cophenetic correlation coefficient results of the linkage methods.

Method	Cophenetic Correlation Coefficient
Single linkage	0.6513
Complete linkage	0.7415
Average linkage	0.7650
Weighted linkage	0.7638
Median linkage	0.6554
Ward linkage	0.7005

Clustering analysis

The purpose of this cluster analysis was to answer RQ1. To answer this question, experiments were carried out on the first three highest clusters with sorted support values. Priorly, we introduce the following terms:

- Top-1 Cluster: Cluster with the first highest support.
- Top-2 Cluster: Cluster with the second highest support.

Table 3: The Davis-Bouldin score of the different cutoff settings.

Cutoff	Davis-Bouldin Score	Number of Clusters
2.5	1.0285	30
2.6	1.0247	29
2.7	1.1144	22
2.8	1.1596	19
2.9	1.2666	15
3.0	1.2774	14
3.1	1.1679	11
3.2	1.2205	9
3.3	1.1790	7
3.4	1.2615	6
3.5	1.2615	6

Table 4: The first three highest clusters and combined clusters among them.

Cluster Name	Disease Member List
Top-1 Cluster	sars cov-2, sars coronavirus (sars-cov) infection, sars-cov- 2, sars-cov-2 virus, sars-cov infection, sars-cov), sars- cov-2 infection, sars-cov-2, sars, sars-cov-2-, cov-2 infection, sars-cov, sars-cov 2, coronavirus sars-cov-2, sars- cov-2, sars-cov-2 coronavirus infection, sars-cov-2 virus infection, sars cov, sars-cov-2 infection
Top-2 Cluster	covid-19-related inflammation, covid-19, covid-19 epidemics, covid-19/, pandemic covid-19, coronavirus pneumonia covid-19, covid-19-associated fever, covid-19 infections, covid-19 pneumonia, covid-19 virus infection, covid-19-related deaths, covid-19 or, covid-19 diseases, covid-19 pulmonary infections, covid-19 global crisis, covid-19 complications, covid-19 pandemics, covid-19 pandemic, covid-19-related cytokine storm, covid-19-associated thrombosis, covid-19 symptoms, covid-19 fatality, covid-19 virus, covid, covid-19 infection
Top-3 Cluster	viremic hiv-1 infection, hiv/sars pseudovirus infection, hiv-1-infected, hiv-1 associated, hiv-1 clade c, hiv-hepatitis c virus, hiv-lipodystrophy, hiv-ltb, hiv-related fatigue, hiv-1 clade c infection, hiv co-infection, hiv illness, hiv-hcv coinfected, hiv lipoatrophy, hiv-1-positive, hiv-1 infected, hcv-hiv, hiv-1c, hiv/hcv co-infection, hiv-1 iiib, hiv coinfection, hiv-negative, hiv-tb, vif-deficient hiv-1, hiv/aids, advanced hiv infection, hiv-associated nephropathy, hcv/hiv-1 co-, hcv/hiv-1 co-infected, hiv-tb coinfection, hiv-positive, hiv-tb iris, eumenorrheic hiv-positive, hiv infection, hiv-1 infection, hiv/hcv coinfected, hiv-infection, hiv-associated neurocognitive disorder, chronic hiv infection, hiv-1 seropositive, hcv-hiv coinfection, hiv-1c infection, hiv/hcv, hiv-infected, hiv-related illness

- Top-3 Cluster: Cluster with the third highest support.
- Top-12 Cluster: Combined Top-1 and Top-2 Clusters
- Top-123 Cluster: Combined Top-1, Top-2 and Top-3 Clusters

The cluster names along with a list of disease as the cluster members are listed in Table 4.

As can be seen in Figure 2, the Top-1 cluster dominated the percentage of COVID-19 drug findings for the determined minimum support. In 8 of 10 minimum support settings, the Top-1 clusters outperformed or were at least equal to the Top-12 clusters. Additionally, in 9 of 10 minimum support settings, the Top-1 clusters outperformed the Top-123 clusters. It was shown that by using the rules in the Top-1 cluster, drug candidates can be found with a higher percentage of findings.

These results indicated that with the Top-1 cluster, drug candidates can be found with a high percentage of findings. In addition, it is computationally beneficial because only using the rules on the Top-1 cluster will save processing time in finding the drug candidates. Table 5 shows that the Top-1 cluster utilized the smallest mining time. The time is calculated from a series of processes in the proposed method, including when connected to the PubChem and DrugBank databases.

From the first three highest support clusters, it was found that the rules in the Top-123 cluster had the lowest performance of findings, while Top-1 and Top-12 are quite similar (see Figure 2). From Table 4, the diseases in the rules of Top-3 cluster are diseases that are dominated by the word "hiv." While in the Top-1 and Top-2, they are diseases with the predominance of the words "sars," "cov," "coronavirus," "covid," and "infection." In this case, we found that those disease keywords have a big influence on drug candidate discovery. The dominance of words can be seen in Figure 3, which shows the wordcloud of each cluster.

Comparison with and without clustering

Text mining with clustering is what we proposed, while text mining without clustering is a comparison that is tested by passing the clustering stage. Thus, the process without clustering meant that from the generating rules stage, it went

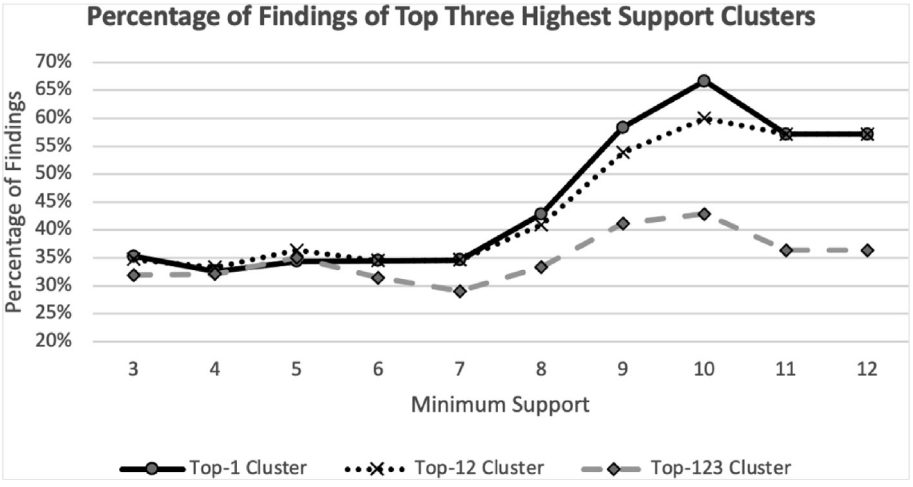

Figure 2: The percentage of COVID-19 drug candidate findings of the top three highest support clusters.

Table 5: Processing time of the first highest clusters and combined clusters with the second and the third ones.

Cluster	Run1	Run2	Run3	Run4	Run5	Mean	StdDev
Top-1	105,80	103,30	104,70	105,70	103,90	104,68	1,10
Top-12	126,60	123,00	123,40	126,00	128,30	125,46	2,23
Top-123	153,40	153,60	152,60	151,90	150,60	152,42	1,22

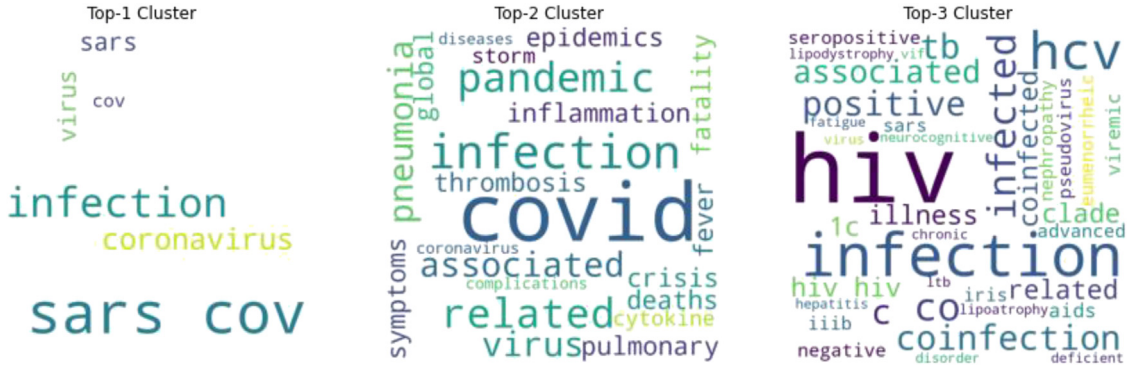

Figure 3: The wordcloud of the first three highest support clusters.

Table 6: The comparison result of text mining with and without the clustering stage.

Min Supp	Without Clustering				With Clustering			
	#Rules	#PubChem	#DrugBank	PF (%)	#Rules	#PubChem	# DrugBank	PF (%)
3	190	179	50	27.93	69	68	24	35.29
4	131	122	37	30.33	45	43	14	32.56
5	93	83	31	37.35	34	32	11	34.38
6	68	60	23	38.33	29	29	10	34.48
7	55	49	18	36.73	26	26	9	34.62
8	51	46	18	39.13	21	21	9	42.86
9	51	46	18	39.13	14	12	7	58.33
10	46	42	17	40.48	11	9	6	66.67
11	42	40	16	40.00	9	7	4	57.14
12	39	37	13	35.14	9	7	4	57.14
Mean				**36.46**				**45.35**

Table 7: Binding energies of 23 medicinal compounds of findings to the four protein structures of SARS-CoV-2.

SMILES	Generic name	Binding energy (kcal/mol)				
		6VW1	6LXT	6VXX	6VYB	Mean
CCC1C(C(C(N(CC(CC(C(C(C(C(C(=O)O1)C)OC2CC(C(C(O2)C)O)(C)OC)C)OC3C(C(CC(O3)C)N(C)C)O)(C)O)C)C)O)(C)O	azithromycin	−8.0	−7.4	−7.9	−7.6	−7.73
COC1=C(C2=C[N+]3=C(C=C2C=C1)C4=CC5=C(C=C4CC3)OCO5)OC	berberine	−7.5	−7.9	−7.8	−8.5	−7.93
CN(C)C(=O)COC(=O)CC1=CC=C(C=C1)OC(=O)C2=CC=C(C=C2)N=C(N)N	camostat	−8.2	−7.7	−8.6	−7.1	−7.90
CC(CS)C(=O)N1CCCC1C(=O)O	captopril	−5.1	−4.6	−5.3	−5.5	−5.13
	chloroquine	−6.3	−6.0	−6.3	−6.5	−6.28

Table 7 (*continued*)

SMILES	Generic name	Binding energy (kcal/mol)				
		6VW1	6LXT	6VXX	6VYB	Mean
CCN(CC)CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl						
COC1=C(C=CC(=C1)C=CC(=O)CC(=O)C=CC2=CC(=C(C=C2)O)OC)O	curcumin	−7.0	−7.2	−6.9	−7.0	−7.03
CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)CO)O)C)O)F)C	**dexamethasone**	**−9.2**	**−8.3**	**−9.0**	**−9.5**	**−9.00**
C1=C(N=C(C(=O)N1)C(=O)N)F	favipiravir	−5.2	−5.9	−5.8	−6.1	−5.75
CC(C)CC(C(=O)NC(CC1CCNC1=O)C(O)S(=O)(=O)O)NC(=O)OCC2=CC=CC=C2	gc376	−7.5	−7.2	−8.2	−7.8	−7.68
CC1C(C(C(C(O1)OCC2C(C(C(C(O2)OC3=CC(=C4C(=O)CC(OC4=C3)C5=CC(=C(C=C5)OC)O)O)O)O)O)O)O)O	**hesperidin**	**−9.4**	**−9.6**	**−9.5**	**−10.2**	**−9.68**
CCN(CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl)CCO	hydroxychloroquine; hcq	−6.4	−6.2	−6.3	−6.4	−6.33
CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	ibuprofen	−6.6	−6.1	−6.7	−6.4	−6.45
CC1=C(C(=CC=C1)C)OCC(=O)NC(CC2=CC=CC=C2)C(CC(CC3=CC=CC=C3)NC(=O)C(C(C)C)N4CCCNC4=O)O	lopinavir	−9.0	−7.8	−9.1	−7.7	−8.40
CCCCC1=NC(=C(N1CC2=CC=C(C=C2)C3=CC=CC=C3C4=NNN=N4)CO)Cl	losartan	−7.6	−7.5	−7.8	−8.0	−7.73
CC(=O)NCCC1=CNC2=C1C=C(C=C2)OC	melatonin	−6.6	−6.2	−6.3	−6.4	−6.38
C1=CC(=CC=C1C(=O)OC2=CC3=C(C=C2)C=C(C=C3)C(=N)N)N=C(N)N	nafamostat	−8.6	−8.0	−9.4	−8.7	−8.68
C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	piperine	−7.6	−7.1	−8.0	−7.9	−7.65
C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O	quercetin	−8.1	−7.7	−8.6	−8.4	−8.20
CC1CCC2CC(C(=CC=CC=CC(CC(C(=O)C(C(C(=CC(C(=O)CC(OC(=O)C3CCCCN3C(=O)C(=O)C1(O2)O)C(C)CC4CCC(C(C4)OC)O)C)C)O)OC)C)C)OC	**rapamycin (sirolimus)**	**−9.4**	**−10.0**	**−9.7**	**−9.1**	**−9.55**
CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4	remdesivir	−8.3	−7.4	−7.3	−8.0	−7.75
C1=CC(=CC=C1C=CC2=CC(=CC(=C2)O)O)O	resveratrol	−7.1	−6.9	−8.0	−7.3	−7.33
C1=NC(=NN1C2C(C(C(O2)CO)O)O)C(=O)N	ribavirin	−6.1	−6.0	−6.7	−6.4	−6.30
CC(C)C1=NC(=CS1)CN(C)C(=O)NC(C(C)C)C(=O)NC(CC2=CC=CC=C2)CC(C(CC3=CC=CC=C3)NC(=O)OCC4=CN=CS4)O	ritonavir	−8.1	−6.9	−8.5	−8.6	−8.03

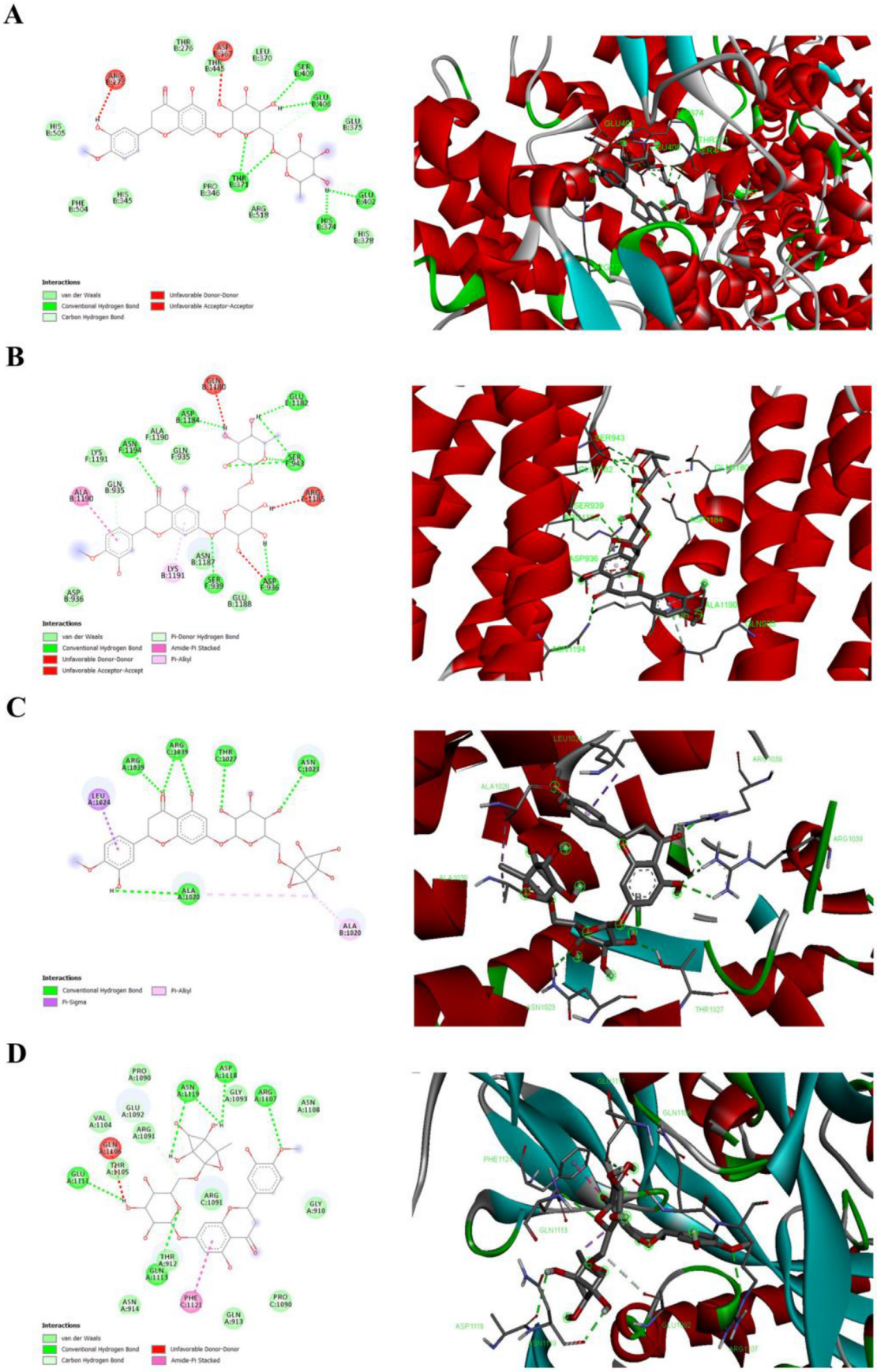

Figure 4: Two-dimensional and three-dimensional molecular interactions of hesperidin against SARS-CoV-2 protein structures. (A) RBD−ACE2 complex (PDB ID: 6VW1). (B) Fusion core of S2 domain (6LXT). (C) Closed state of spike ectodomain (PDB ID: 6VXX). (D) Open state of spike ectodomain (PDB ID: 6VYB).

Table 8: The results of the number of trials, number of trials has results and phases for the 23 predicted drugs.

Generic name	Number of trials	Number of trials (have results)	Phases
hesperidin	1	1	Phase 2
rapamycin (sirolimus)[a]	5 (4)	0 (0)	
dexamethasone	42	2	Phase 3; not available
nafamostat	7	0	
lopinavir	28	0	
quercetin	6	0	
ritonavir	46	0	
berberine	1	0	
camostat	18	4	Phase 2; Phase 2; Phase 2; Phase 2
remdesivir	61	9	Phase 3; Phase 3; Phase 3; Phase 3; Phase 1|Phase 2; Phase 3; Phase 3; Phase 3; Phase 3
azithromycin	50	6	Phase 2; Phase 3; Phase 3; Phase 2; Phase 2; Phase 2
losartan	10	3	Phase 2; Phase 2; Phase 4
gc376	0	0	
piperine	0	0	
resveratrol	3	0	
curcumin	3	0	
ibuprofen	2	0	
melatonin	8	0	
Hydroxychloroquine (hcq)[a]	188 (7)	188 (7)	Phase 2; Phase 2; Phase 2; Phase 2; Phase 3; Phase 2; Phase 2; Phase 2; Phase 1| Phase 2; Phase 2; Phase 2|Phase 3; Phase 3; Phase 2; Phase 2|Phase 3; Phase 3; Phase 2; Phase 3; Phase 2; Phase 3; Phase 3 (Phase 2; Phase 3; Phase 2; Phase 2; Phase 2; Phase 3; Phase 2)
ribavirin	7	1	Phase 2
chloroquine	19	0	
favipiravir	43	3	Phase 3; Phase 2; Phase 2
captopril	2	0	

[a] Denotes synonyms names.

directly to the extracting drugs stage of our proposed method. This comparison is intended to determine the effect of clustering on the proposed mining process to find drug candidates as well as to answer RQ2. The results of their comparison can be found in Table 6. It suggested that the number of rules generated by mining with clustering obtains a better percentage of findings on average for all minimum support settings. In 7 of 10 minimum support settings (PF with underlining text in Table 6), the mining with clustering outperformed the mining without clustering; hence, the method is quite promising in identifying drug candidates through the search of biomedical literature.

Discussion

Natural compound utilization is not a new idea in the struggle to combat COVID-19. There are at least two published reports documenting a total of 38 ongoing clinical trials of herbal medicines for the treatment of COVID-19.[20,33] This study aimed to obtain a list of COVID-19 drug candidates through the disease-drug associations extraction approach from the search of related literature. This study applied the AHC method to obtain a collection of drug candidates that meet a certain number of disease-drug associations frequencies. We applied the minimum support values to test the effect of the frequency of associations on the percentage of drug candidates recorded in PubChem and DrugBank.

It was found that the higher the minimum support, the less the number of rules obtained, and this has an impact on the number of findings (see Table 6). While the minimum support was equal to 3, 68 compounds were found in PubChem and 24 of them were found to be related to COVID-19 according to the COVID-19 DrugBank. While the minimum support was equal to 12, only 7 and 4 compounds were found in PubChem and DrugBank, respectively. From Figure 2, it can be seen that when the minimum support was increased, the percentage of findings tended to increase; however, the number of drug compounds found have decreased. Although minimum support equal to 12 obtained higher percentage than minimum support equal to 3, the number of drug candidates found was fewer. Therefore, for identifying the more of the COVID-19 drug candidates, we suggested that the determination of this minimum support was equal to 3 and identified 24 drug

candidates. The 24 drug candidates involved 23 unique SMILES formulas as a standard representation of chemical structures.

To demonstrate the feasibility and validity of our proposed approach, there were 23 drug candidates in which we estimated the binding energy by using the drug docking calculation. The binding energy was used as the evaluation indicator to judge the binding affinity of drug candidates against SARS-CoV-2 protein targets. This validation method was similar to that of previously published studies of virtual screening of COVID-19 drugs. A prior study conducted a simulated molecular docking experiments with several natural compounds from propolis extract, and then observed the higher binding affinity of methylophiopogonone A, 3′-methoxydaidzin, and genistin to two protein structures of SARS-CoV-2 including main protease protein and spike protein subunit 2.[31] Another study speculated that a natural compound, 4-gingerol, has good binding affinity against SARS-CoV-2 main protease by performing a molecular docking experiment.[34] In our study, the AutoDock Vina runs resulted in the average values of binding energy from −5.13 to −9.68 kcal/mol to the four SARS-CoV-2 protein structures viz. 6VW1, 6LXT, 6VXX, and 6VYB (Table 7). According to the estimated binding affinity, three drug candidates were further identified with the lowest average values of binding energy, including hesperidin (binding energy, −9.68 kcal/mol), rapamycin or sirolimus (binding energy, −9.55 kcal/mol), and dexamethasone (binding energy, −9.00 kcal/mol).

From the PubMed searching, a total of 26, 10, and 258 publications were retrieved based on the disease keywords in the Top-1 cluster coupled with the drug/phytochemical name of hesperidin, sirolimus, and dexamethasone, respectively. Among these articles, there have been at least three articles reported that hesperidin could be used to prevent SARS-CoV-2 infection.[35–37] Moreover, at least two articles reported that hesperidin has the potential to inhibit the SARS-CoV-2 virus entry by blocking the binding of the virus to the ACE2 receptor protein.[38,39] One of the 26 articles showed that hesperidin had better binding affinity than nelfinavir, chloroquine, and hydroxychloroquine as spike glycoprotein inhibitors.[40] Another study by Utomo et al. also conducted a docking analysis to elucidate the potential of hesperidin in binding the SARS-CoV-2 protease, spike protein, transmembrane serine protease 2, and PD-ACE2 with higher binding affinity compared to several existing viral drugs such as lopinavir, nafamostat, and comastat.[41] These results are consistent with our results of drug docking in Table 7. A total of six, eight, six, and six conventional hydrogen bonds found from four SARS-CoV-2 protein structures, including the RBD−ACE2 complex, the fusion core of S2 domain, and the closed state and open state of spike ectodomain (Figure 4). Hydrogen binding can provide a stabilization effect and determine the key interacted residues. Therefore, a high number of hydrogen bonds that were formed may be associated with the highest affinity between hesperidin and SARS-CoV-2 protein.

From these searched articles, there are a lack of articles to report the mechanism of action against SARS-CoV2 protein among the 10 articles about the COVID-19 treatment with sirolimus. However, most of these articles point to sirolimus as an immunosuppressive drug and thus could be as a potential medicine for the treatment of immunocompromised status in COVID-19 patients.[42,43] Our docking results also confirmed that sirolimus has the second highest affinity against SARS-CoV-2 spike proteins.

Several publications reported the usage of dexamethasone to combat COVID-19 with positive results, such as reducing the death rate of patients with and without mechanical ventilation,[44] having inhibitory activities against COVID-19 proteases,[45] regulating ACE2,[46,47] and having modest effects in moderate and severe COVID-19.[48] However, some studies found that the methylprednisolone[49] and oral prednisone have better effects than dexamethasone on recovery time, intensive care needs, and the level of severity biomarkers.[50] Another study did not suggest that dexamethasone be routinely prescribed for COVID-19 patients after discharge.[51] It is also notable that corticosteroids like dexamethasone should be carefully used because of the risk that it carries.[45] Nevertheless, our docking results were in line with the effect of dexamethasone to COVID-19, and also illustrated the detailed bonding of molecular interaction in the Appendix (Figures A.1 and A.2). By reviewing these articles and referring to the drug docking results, we also demonstrated the validity of disease keywords in the Top-1 cluster and clarified the mechanism of action of the top three drug candidates.

Furthermore, we utilized clinical trial data to validate our predicted drugs. We used the listed clinical studies related to COVID-19, ClinicalTrials.gov (https://clinicaltrials.gov/), and obtained a total of 8108 entries (August 5, 2022 version). Then we compared the 23 drugs with those entries, and the results are listed in Table 8. A total of 21 drugs (91.3% of our prediction), except gc376 and piperine, have clinical trials. In Table 8, we present the number of trials, number of trials that have results and phases as reported by ClinicalTrials.gov. More detailed information is given in Supplementary file 1.

Also, since there may be concerns about the possibility of risk of bias in our studies, we employed the following steps to reduce the risk of bias and enhance the significance of our findings.

[1] Clustering stage

We have TF-IDF as the features and the employed AHC to group similar diseases and then identified the associations between COVID-19 and drugs found in the abstract of the articles.

[2] Evaluation metrics

We used the cophenetic correlation coefficient to determine the best performance of the AHC linkage methods. Furthermore, we used the Davis-Bouldin score to select the best performance of cluster formation. Both methods are implemented to reduce bias in our study.

[3] Ranking of potential drugs

We performed binding energy calculations to rank the 23 potential drugs. The calculation considered four protein

domain structures of SARS-CoV-2 and used the average binding energy value as the basis for ranking.

Limitations and future work

We only used PubMed as the source of scientific literature data. There are other databases that contain scientific literature, such as Medline, TOXLINE, Embase, BIOSIS Citation Index (Web of Science), Biological Sciences (ProQuest), and SciFinder-n. Moreover, we used AHC methods, as well as the partitional clustering method is another approach that can be used. We plan to include other biomedical literature databases while evaluating various clustering methods for optimal results. In addition, we will also develop a web-based application that implements our proposed method as a portal of information in the discovery of drug candidates for COVID-19.

Conclusion

In this study, we applied a novel approach involving BioBERT mining, the usage of agglomerative hierarchical clustering, and utilizing the PubChem database and DrugBank to identify potential COVID-19 drug candidates through disease-drug association rules. The use of the clustering method in mining has been shown to have a positive impact on the discovery of COVID-19 drug candidates. Twenty-four out of sixty-eight drug compounds are confirmed COVID-19 drug candidates according to DrugBank. Finally, we found that hesperidin has the lowest binding energy with the spike protein of the COVID-19 disease. The results have also been confirmed by at least three recent publications. We believe this knowledge base will help the research community explore the existing drugs and biomedical entities for coronavirus-related diseases and find effective treatments for COVID-19.

Availability of data and materials

The datasets analyzed for the current study are publicly available from the NCBI PubMed database.

Source of funding

AAS, VZ, RSY, AA, and HFP works are supported by Research Program (DIPA Rumah Program) at the Research Organization for Life Sciences and Environment, the National Research and Innovation Agency (BRIN) (grant number: 9/III/HK/2022). CWW work is supported by the Taiwan Ministry of Science and Technology (MOST) (grant number: MOST 108-2314-B040-034-MY3). CHH by MOST (grant number: MOST 109-2221-E-150-036). KLN work is supported by MOST (grant numbers: MOST 109-2221-E-468-013). KLN work is also supported by Asia University (grant number: ASIA-110-CMUH-12).

Conflict of interest

The authors have no conflict of interest to declare.

Ethical approval

Not applicable – no human subjects or animals are involved in this study. Our submitted work does not contain contents that have been published or are under consideration for publication by any other journals or conferences.

Authors contributions

AAS, RSY and CWW analyzed and interpreted data; data validation; wrote initial and final draft of article; reviewed the article. VZ, RSY and AA provided data curation and investigation. HFP investigated the data and reviewed the article. CHH provided funding acquisition and reviewed the article. KLN conceived and designed the study; provided funding acquisition; project administration and provided logistical support; wrote initial and final draft of article; reviewed the article. All authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jtumed.2022.12.015.

References

1. Hatmi ZN. A systematic review of systematic reviews on the COVID-19 pandemic. **SN Compr Clin Med 2021**; 3(2): 419–436.
2. Rismanbaf A. Potential treatments for COVID-19; a narrative literature review. **Arch Acad Emerg Med 2020**; 8(1): e29.
3. Chiu YJ, Chiang JH, Fu CW, Hour MJ, Ha HA, Kuo SC, et al. Analysis of COVID-19 prevention and treatment in Taiwan. **Biomedicine (Taipei) 2021**; 11(1): 1–18.
4. Zeydi AE, Ghazanfari MJ, Sanandaj FS, Panahi R, Mortazavi H, Karimifar K, et al. Coronavirus Disease 2019 (COVID-19): a literature review from a nursing perspective. **Biomedicine (Taipei) 2021**; 11(3): 5–14.
5. Kuusisto F, Page D, Stewart R. Word embedding mining for SARS-CoV-2 and COVID-19 drug repurposing. **F1000Research 2020**; 9(585).
6. Muramatsu T, Tanokura M. A novel method of literature mining to identify candidate COVID-19 drugs. **Bioinform Adv 2021**; 1(1): 1–5.
7. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. **J Am Med Inf Assoc 2008**; 15(1): 87–98.
8. 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(11):e21648.
9. Wang J, Wang W, Yan C, Luo J, Zhang G. Predicting drug-disease association based on ensemble strategy. **Front Genet 2021**; 12:666575.
10. Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB, Olsen LR. BioReader: a text mining tool for performing classification of biomedical literature. **BMC Bioinf 2019**; 19(Suppl 13): 57.
11. Zhou J, Fu BQ. The research on gene-disease association based on text-mining of PubMed. **BMC Bioinf 2018**; 19(1): 37.

12. Chaix E, Deléger L, Bossy R, Nédellec C. Text mining tools for extracting information about microbial biodiversity in food. **Food Microbiol 2019**; 81: 63–75.
13. Macnee M, Pérez-Palma E, Schumacher-Bass S, Dalton J, Leu C, Blankenberg D, et al. SimText: a text mining framework for interactive analysis and visualization of similarities among biomedical entities. **Bioinformatics 2021**: 1–3.
14. Geoffrey ASB, Preetham P, Sanker A, Madaj R, Antony H. *Compound2Drug - a machine/deep learning tool for predicting the bioactivity of PubChem compounds*. ChemRxiv. Biological and Medical Chemistry. Cambridge: Cambridge Open Engage; 2020.
15. Zaslavsky L, Cheng T, Gindulyte A, He S, Kim S, Li Q, et al. Discovering and summarizing relationships between chemicals, genes, proteins, and diseases in PubChem. **Front Res Metr Analy 2021**; 6:689059.
16. Barupal DK, Fiehn O. Generating the blood exposome database using a comprehensive text mining and database fusion approach. **Environ Health Perspect 2019**; 127(9):97008.
17. Papanikolaou N, Pavlopoulos GA, Theodosiou T, Vizirianakis IS, Iliopoulos I. DrugQuest - a text mining workflow for drug association discovery. **BMC Bioinf 2016**; 17(Suppl 5): 182.
18. Ben Abacha A, Chowdhury MFM, Karanasiou A, Mrabet Y, Lavelli A, Zweigenbaum P. Text mining for pharmacovigilance: using machine learning for drug name recognition and drug-drug interaction extraction and classification. **J Biomed Inf 2015**; 58: 122–132.
19. Dere S, Ayvaz S. Prediction of drug–drug interactions by using profile fingerprint vectors and protein similarities. **Healthc Inform Res 2020**; 26(1): 42–49.
20. Yang Y, Islam MS, Wang J, Li Y, Chen X. Traditional Chinese medicine in the treatment of patients infected with 2019-new coronavirus (SARS-CoV-2): a review and perspective. **Int J Biol Sci 2020**; 16(10): 1708–1717.
21. Huang YF, Bai C, He F, Xie Y, Zhou H. Review on the potential action mechanisms of Chinese medicines in treating Coronavirus Disease 2019 (COVID-19). **Pharmacol Res 2020**; 158:104939.
22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics 2020**; 36(4): 1234–1240.
23. Giles O, Karlsson A, Masiala S, White S, Cesareni G, Perfetto L, et al. Optimising biomedical relationship extraction with BioBERT. **bioRxiv 2020**. https://doi.org/10.1101/2020.09.01.277277.
24. Zhu Y, Li L, Lu H, Zhou A, Qin X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. **J Biomed Inf 2020**; 106:103451.
25. Weininger D. SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules. **J Chem Inf Comput Sci 1988**; 28(1): 31–36.
26. Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. **Taxon 1962**; 11(2): 33–40.
27. Davies DL, Bouldin DW. A cluster separation measure. **IEEE Trans Pattern Anal Mach Intell 1979**; PAMI-1(2): 224–227.
28. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. **Nature 2020**; 581(7807): 221–224.
29. Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. **Cell Res 2020**; 30(4): 343–355.
30. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. **Cell 2020**; 181(2): 281–292 e6.
31. Harisna AH, Nurdiansyah R, Syaifie PH, Nugroho DW, Saputro KE, Firdayani, et al. In silico investigation of potential inhibitors to main protease and spike protein of SARS-CoV-2 in propolis. **Biochem Biophys Rep 2021**; 26:100969.
32. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **J Comput Chem 2010**; 31(2): 455–461.
33. Lee DYW, Li QY, Liu J, Efferth T. Traditional Chinese herbal medicine at the forefront battle against COVID-19: clinical experience and scientific basis. **Phytomedicine 2021**; 80: 153337.
34. Wijaya RM, Hafidzhah MA, Kharisma VD, Ansori ANM, Parikesit AA. COVID-19 in silico drug with zingiber officinale natural product compound library targeting the Mpro protein. **Makara J Sci 2021**; 25(3): 162–171.
35. Basu A, Sarkar A, Maulik U. Molecular docking study of potential phytochemicals and their effects on the complex of SARS-CoV2 spike protein and human ACE2. **Sci Rep 2020**; 10(1):17699.
36. Bellavite P, Donzelli A. Hesperidin and SARS-CoV-2: new light on the healthy function of citrus fruits. **Antioxidants 2020**; 9(8).
37. Cheng FJ, Huynh TK, Yang CS, Hu DW, Shen YC, Tu CY, et al. Hesperidin is a potential inhibitor against SARS-CoV-2 infection. **Nutrients 2021**; 13(8).
38. Haggag YA, El-Ashmawy NE, Okasha KM. Is hesperidin essential for prophylaxis and treatment of COVID-19 infection? **Med Hypotheses 2020**; 144:109957.
39. Joshi RS, Jagdale SS, Bansode SB, Shankar SS, Tellis MB, Pandya VK, et al. Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. **J Biomol Struct Dyn 2021**; 39(9): 3099–3114.
40. Tallei TE, Tumilaar SG, Niode NJ, Fatimawali, Kepel BJ, Idroes R, et al. Potential of plant bioactive compounds as SARS-CoV-2 main protease (M(pro)) and spike (S) glycoprotein inhibitors: a molecular docking study. **Scientifica (Cairo) 2020**; 2020:6307457.
41. Utomo RY, Ikawati M, Putri DDP, Salsabila IA, Meiyanto E. The chemopreventive potential of diosmin and hesperidin for COVID-19 and its comorbid diseases. **Indones J Cancer Chemoprevention 2020**; 11(3): 154–167.
42. Manzo ML, Galati C, Gallo C, Santangelo G, Marino A, Guccione F, et al. ADEM post-Sars-CoV-2 infection in a pediatric patient with Fisher-Evans syndrome. **Neurol Sci 2021**; 42(10): 4293–4296.
43. Talwar D, Kumar S, Acharya S, Hulkoti V, Annadatha A. Sirolimus in a renal transplant recipient infected with COVID-19: a blessing in disguise? **Cureus 2021**; 13(8):e17102.
44. Andreakos E, Papadaki M, Serhan CN. Dexamethasone, pro-resolving lipid mediators and resolution of inflammation in COVID-19. **Allergy 2021**; 76(3): 626–628.
45. Noreen S, Maqbool I, Madni A. Dexamethasone: therapeutic potential, risks, and future projection during COVID-19 pandemic. **Eur J Pharmacol 2021**; 894:173854.
46. Sinha S, Cheng K, Schaffer AA, Aldape K, Schiff E, Ruppin E. In vitro and in vivo identification of clinically approved drugs that modify ACE2 expression. **Mol Syst Biol 2020**; 16(7):e9628.
47. Saheb Sharif-Askari N, Saheb Sharif-Askari F, Alabed M, Tayoun AA, Loney T, Uddin M, et al. Effect of common medications on the expression of SARS-CoV-2 entry receptors in kidney tissue. **Clin Transl Sci 2020**; 13(6): 1048–1054.
48. Asselah T, Durantel D, Pasmant E, Lau G, Schinazi RF. COVID-19: discovery, diagnostics and drug development. **J Hepatol 2021**; 74(1): 168–184.
49. Ko JJ, Wu C, Mehta N, Wald-Dickler N, Yang W, Qiao R. A comparison of methylprednisolone and dexamethasone in

intensive care patients with COVID-19. **J Intensive Care Med 2021**; 36(6): 673–680.
50. Pinzon MA, Ortiz S, Holguin H, Betancur JF, Cardona Arango D, Laniado H, et al. Dexamethasone vs methylprednisolone high dose for Covid-19 pneumonia. **PLoS One 2021**; 16(5):e0252057.
51. Huang CW, Yu AS, Song H, Park JS, Wu SS, Khang VK, et al. Association between dexamethasone treatment after hospital discharge for patients with COVID-19 infection and rates of hospital readmission and mortality. **JAMA Netw Open 2022**; 5(3):e221455.

How to cite this article: Supianto AA, Nurdiansyah R, Weng C-W, Zilvan V, Yuwana RS, Arisal A, Pardede HF, Lee M-M, Huang C-H, Ng K-L. Cluster-based text mining for extracting drug candidates for the prevention of COVID-19 from the biomedical literature. J Taibah Univ Med Sc 2023;18(4):787–801.