Text and Network-Mining for COVID-19 Intervention Studies

Aditya Rao (adityar.rao@tcs.com) Saipradeep VG (saipradeep.v@tcs.com)
Thomas Joseph (thomas.joseph@tcs.com) Sujatha Kotte (kotte.sujatha@tcs.com)
Naveen Sivadasan (naveen.sivadasan@tcs.com)
Rajgopal Srinivasan* (rajgopal.srinivasan@tcs.com)

TCS Research and Innovation, Hyderabad-500081, INDIA

Abstract

Background: *The COVID-19 pandemic has led to a massive and collective pursuit by the research community to find effective diagnostics, drugs and vaccines The large and growing body of literature present in MEDLINE and other online resources including various self-archive sites are invaluable for these efforts. MEDLINE has more than 30 million abstracts and an additional corpus related to COVID-19, SARS and MERS has more than 40,000 literature articles, and these numbers are growing. Automated extraction of useful information from literature and automated generation of novel insights is crucial for accelerated discovery of drug/vaccine targets and re-purposing drug candidates.*

Methods: *We applied text-mining on MEDLINE abstracts and the CORD-19 corpus to extract a rich set of pair-wise correlations between various biomedical entities. We built a comprehensive pair-wise entity association network involving 15 different entity types using both text-mined associations as well as novel associations obtained using link prediction. The resulting network, which we call CoNetz, also contains a specialized COVID-19 subnetwork that provides a network view of COVID-19 related literature. Additionally, we developed a set of network exploration utilities and user-friendly network visualization utilities using NetworkX and PyVis.*

Results: *CoNetz consisted of pair-wise associations involving ∼174,000 entities covering 15 different entity types. The specialized COVID-19 subnetwork consisted of ∼7.8 million pair-wise associations involving ∼43,000 entities. The network captured several of the well-known COVID-19 drug re-purposing candidates and also predicted novel candidates including ingavirin, laninamivir, nevirapine, paritaprevir, pranlukast and peficitinib.*

Conclusions: *Our automated text and network-mining approach builds an up-to-date and comprehensive knowledge network from literature for COVID-19 studies. The wide range of entity types captured in CoNetz provides a rich neighborhood context around the relations of interest. The approach avoids multiple drawbacks associated with manual curation including cost and effort involved, lack of up-to-date information and limited coverage. Amongst the novel repurposing drugs predicted, laninamivir and paritaprevir are possible COVID-19 anti-viral drugs while pranlukast was postulated to be a candidate for managing severe respiratory symptoms in COVID-19 patients. CoNetz is available for download and use from https://web.rniapps.net/tcn/tcn.tar.gz*

Keywords: *COVID-19, SARS-CoV-2, biomedical text-Mining, MEDLINE, TPX, HANRD, GCAS, Graph Convolution, link-prediction*

1 Background

With over 2.5 million cases worldwide in less than four months, the coronavirus disease 2019 (COVID-19) pandemic has caught the administration, health experts and the general public by surprise [1]. COVID-19 is caused by a novel betacoronavirus, now named SARS-CoV-2. SARS-CoV-2 shares 79% sequence identity with SARS-CoV, the virus which caused a major severe acute respiratory syndrome (SARS) outbreak in 2002-2003 [2]. Yet COVID-19 has proved quite different from SARS and Middle East respiratory syndrome (MERS), both in disease progression, complications and mortality. This has led to a massive and collaborative ongoing pursuit by the research community to find drugs and vaccines for the disease. While efforts to contain the pandemic through non-pharmaceutical interventions have helped contain the spread of COVID-19, there is a an urgent need for a broad-based and effective pharmaceutical intervention. These could be through repurposing or repositioning of existing drugs or vaccines, or by identifying novel ones. The drug repurposing strategy is faster since it involves the use of drugs that have been approved for use in humans, with their pharmacology and toxicity aspects already known. Finding drugs that can be repurposed for COVID-19 remains a very active strategy [3, 4, 5, 6].

Some of the studies used network-based approaches for systematic identification of repurposable drugs and drug combinations for potential treatment of COVID-19 [3, 7]. The network here represents the interactome of human protein-protein interaction pairs. The basis for identifying drug repurposable candidates using such networks is that the proteins that associate with and functionally govern COVID-19 infection are localized in a subnetwork of the human interactome in the neighborhood of SARS-CoV-2 [3, 5].

Curated interactome data suffer from several drawbacks, including time and effort involved, lack of up-to-date information, limited recall, and limited coverage of entity types. A significant fraction of these interactions are derived from literature. The large and growing body of literature present in MEDLINE and other online resources including various self-archive sites are the primary source of up-to-date interactome data. However, approaches requiring any manual intervention to information extraction from literature is highly infeasible due their massive size. For example, MEDLINE has more than 30 million abstracts and an additional corpus related to COVID-19, SARS and MERS has more than 40,000 literature articles [8]. Hence, automated extraction of useful information from literature and generation of novel insights is crucial for rapidly identifying of drug/vaccine targets and re-purposing drug candidates.

We have previously described TPX, a text-mining framework that supports real-time entity assisted search and navigation of MEDLINE and literature corpora [9]. We have also described our link-prediction algorithm GCAS (*Graph Convolution-based Association Scoring*) that is able to predict novel links between entities [10]. In another study, GCAS was applied on a MEDLINE-derived association network for rare disease gene prioritization [11].

In this study, we tailored TPX modules for automated text-mining of literature towards constructing a comprehensive and up-to-date correlation network for COVID-19 studies. We then applied GCAS-based network mining on the network to infer novel links, and augment the correlation network with these inferred links. We illustrated the value of the resulting network by investigating subnetworks with the aim of identifying potential drugs both for the treatment and management of COVID-19.

2 Methods

Figure 1 gives the overall architecture of our text and network-mining based framework. Broadly, our system consists of (a) Dictionary Curation module, (b) Named Entity Recognition (NER) module for corpus annotation (c) Correlation Extraction Module, for identifying correlated entity pairs (d) Link-prediction module to predict novel links and (e) Network exploration and visualization module. We discuss each of these modules in this section.

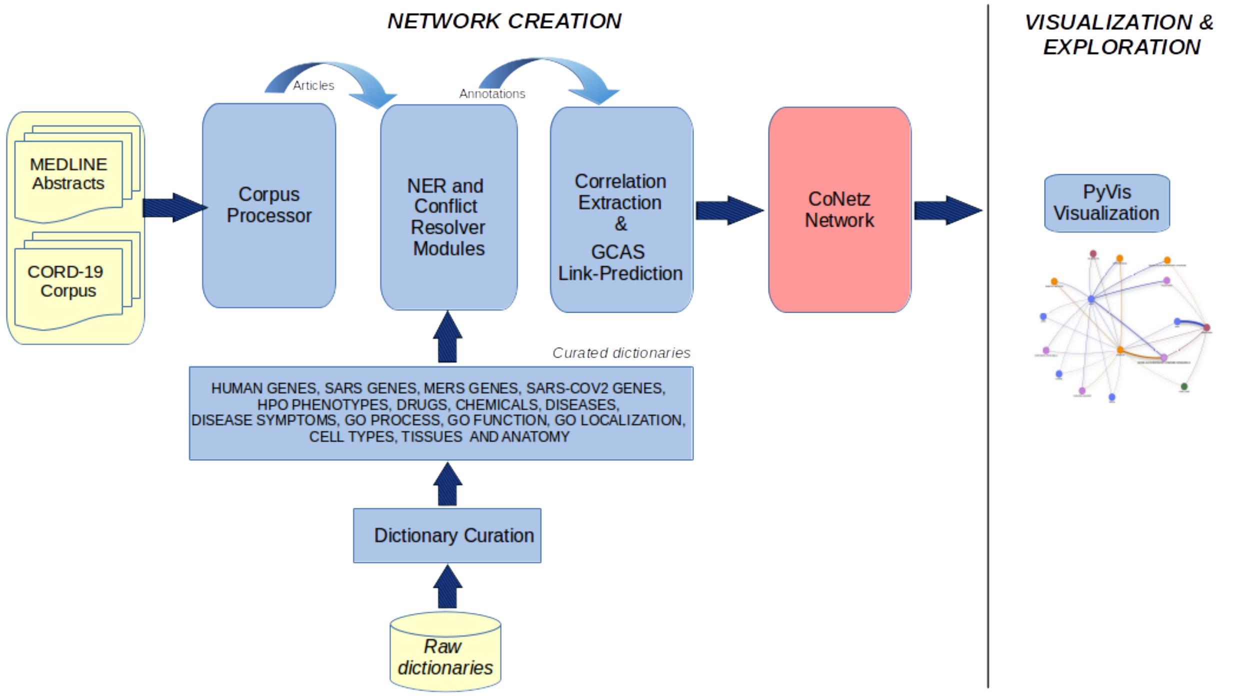
NETWORK CREATION
VISUALIZATION & EXPLORATION
MEDLINE Abstracts
CORD-19 Corpus
Articles
Corpus Processor
Annotations
NER and Conflict Resolver Modules
Correlation Extraction & GCAS Link-Prediction
CoNetz Network
PyVis Visualization
Curated dictionaries
HUMAN GENES, SARS GENES, MERS GENES, SARS-COV2 GENES, HPO PHENOTYPES, DRUGS, CHEMICALS, DISEASES, DISEASE SYMPTOMS, GO PROCESS, GO FUNCTION, GO LOCALIZATION, CELL TYPES, TISSUES AND ANATOMY
Dictionary Curation
Raw dictionaries

Figure 1: Architecture depicting the *precomputation phase* where the network is built, and the *visualization phase* where the visualization and exploration are carried out.

2.1 Corpus

We included all MEDLINE abstracts in the input corpus (Date of download: April 5th, 2020). Additionally, we included articles from the COVID-19 Open Research Dataset (CORD-19 dataset, downloaded in April, 2020) [8]. This is a freely-available growing corpus of abstracts and full-text articles about COVID-19 and related coronavirus infections. This set also contained articles from various self-archive sites.

2.2 Dictionary Curation

We created a comprehensive set of dictionaries of names and synonyms mapping to 15 different entity types including human genes, SARS genes, MERS genes, SARS-CoV-2 genes, HPO phenotypes, drugs, chemicals, diseases, disease symptoms, GO process, GO function, GO localization, cell types, tissues and anatomy. Such a rich set of dictionaries were used so that (a) the associations identified have a wide variety, and (b) the network neighborhood of any identified association provides a rich context. The dictionaries were built from various sources such as MeSH [12], Gene Ontology [13], HGNC [14], amongst others. Table 1 provides details of the entities and their primary sources.

Entity ID	Entity Type	Entity Source
1	HUMAN GENE	https://www.genenames.org/
2	SARS GENE	https://www.ncbi.nlm.nih.gov/gene/
3	MERS GENE	https://www.ncbi.nlm.nih.gov/gene/
4	SARS-CoV-2 GENE	https://www.ncbi.nlm.nih.gov/gene/
5	HPO PHENOTYPES	https://hpo.jax.org/
6	DISEASE	https://www.nlm.nih.gov/mesh/meshhome.html
7	CHEMICALS	https://www.nlm.nih.gov/mesh/meshhome.html
8	DRUGS	https://www.nlm.nih.gov/mesh/meshhome.html
9	DISEASE SYMPTOMS	https://www.nlm.nih.gov/mesh/meshhome.html
10	GO PROCESS	http://amigo.geneontology.org/amigo/landing
11	GO FUNCTION	http://amigo.geneontology.org/amigo/landing
12	GO LOCATION	http://amigo.geneontology.org/amigo/landing
13	CELLTYPE	https://www.nlm.nih.gov/mesh/meshhome.html
14	TISSUE	https://www.nlm.nih.gov/mesh/meshhome.html
15	ANATOMY	https://www.nlm.nih.gov/mesh/meshhome.html

Table 1: Various biomedical entities and their respective sources

As with most sources of such data for dictionary creation, issues of overlaps within and across dictionaries, ambiguity as well as issues of coverage were observed. Table 2 shows some of these conflicts. These were addressed to a significant extent with the semi-automated dictionary curation process. As deemed required, manual disambiguation was also done. Further analyses of the tagged output was used to identify noisy acronyms and high-level terms, which were removed as required. Most other ambiguities were resolved using the Conflict Resolver module (Section 2.3) during the tagging phase. We acknowledge that this is an ongoing process, and the results of repeated iterations of tagging will be used to continuously refine these dictionaries.

2.3 Named Entity Recognition (NER)

We tailored two modules from our TPX text-mining framework - the NER module that annotates the literature from MEDLINE and the CORD-19 corpus using the curated dictionaries; and the Conflict

Overlap Type	**Entity**	**ID1**	**ID2**
HUMAN GENE	cox1 cop1	HGNC:9604 HGNC:17440	HGNC:7419 HGNC:33701
GO FUNCTION	trehalose synthase activity cdk-activating kinase activity	GO:0102986 GO:0019912	GO:0047471 GO:0004693
GO PROCESS	axon cargo transport	GO:0098930	GO:0008088
HUMAN GENE-HPO PHENOTYPES	strabismus supravalvular aortic stenosis	HP:0000486 HP:0004381	HGNC:15511 HGNC:3327
HUMAN GENE-DISEASE	familial hypercholesterolemia gyrate atrophy	HGNC:6547 HGNC:8091	D006938 D015799
HPO PHENOTYPES-GO PROCESS	aggressive behavior skeletal muscle atrophy	HP:0000718 HP:0003202	GO:0002118 GO:0014732

Table 2: Examples of overlaps within and across dictionaries

Resolver (CR) module which carries out entity type disambiguation.

The dictionary-based NER module from TPX was used to tag term mentions in the text using the curated dictionaries from section 2.2. Specialized NER modules were developed to handle each entity type differently. For instance, disease terms in the text are tokenized and matched differently compared to the gene terms. The NER module removes stop words corresponding to each dictionary type using separate entity specific stop word lists. Conflicts across entity types were resolved by the CR module during the tagging phase [9, 11].

2.4 Correlation Network Construction

The correlation extraction module used Pearson correlation coefficient to compute pairwise entity correlations between entities identified by the corpus annotator. The correlation score computation is detailed in [11]. The pairwise correlations were derived based on co-occurrences of entity pairs at paragraph level in text. We experimented with computing pair-wise correlations using different text spans: sentence, paragraph, section and article. We observed that correlations computed at the paragraph level struck a good balance between precision and recall. The resulting network captured entity pairs with positive correlations. Here, the correlation values are stored as link weights.

Additionally, we computed the entity correlations by restricting the corpus to only COVID-19 and related articles. This is to enable users to explore and analyze the pair-wise correlations in the context of COVID-19 and related literature as a specialized COVID-19 subnetwork. The specialized COVID-19 corpus consisted of the CORD-19 corpus and a filtered set of MEDLINE abstracts. The MEDLINE abstracts were filtered based on the annotated entities in the abstracts, wherein the included abstracts contained at least one tagged entity corresponding to “coronavirus infection”, SARS, MERS or

COVID-19 or their synonyms. The pair-wise correlation values derived from this COVID-19 context are included as additional edge attributes in the network. The pair-wise correlation values are governed by both the co-occurrence frequencies of the entity pairs as well as their individual background frequencies in the corpus. Hence, by limiting the corpus to COVID-19 related literature, the entity pairs could have altered correlation values in the specialized subnetwork. This subnetwork serves as a network view of COVID-19 related literature.

2.5 Link Prediction

Drug and vaccine research can greatly benefit from automatic prediction of novel links that in turn helps in improved hypothesis generation. Since the existing literature is an important source for effective link prediction, we applied the GCAS link prediction algorithm [10] to the entire correlation network i.e., correlations from the entire corpus, thereby leveraging the available knowledge from literature to predict novel associations. GCAS identifies novel links by performing information propagation on the network using spectral graph convolution techniques [10]. The final network obtained after running GCAS also included the predicted links. We refer to this network as "CoNetz".

2.6 Network Exploration and Visualization

We built PyVis and NetworkX-based [15] utilities for easy exploration and intuitive visualization of the network. PyVis is a python-based framework for visualization of network graphs based on the VisJS library [16]. PyVis provides a rich set of customizable visualization features including node colors, labels, sizes based on metadata and a wide choice of layout algorithms. User interface operations such as dragging, hovering, and selection of nodes and edges are well supported in PyVis. Here, the nodes represent various biomedical entities, while the edges represent the correlations or links between them. The width of the edge indicates the association score between the nodes. The edges also provide the evidence of an association as a list of article IDs. The supporting literature for correlation links can be viewed as part of the edge information. NetworkX supports querying an underlying network-based on multiple criteria in order to, for example, identify important links or to identify paths that connect distant entity nodes. We used NetworkX to: (a) perform complex queries to explore connecting nodes, k$^{\text{th}}$-neighborhood links between query node(s), and (b) query with a wide range of filters viz., *query by term*, *show only defined entity types*, *display top-K entities across or within each entity type*, amongst others. Thus, an underlying network can be effectively explored using NetworkX-based filtering of the relevant sub-networks and its visualization using PyVis.

3 Results

Our final network, CoNetz, consisted of ∼174,000 unique entities and ∼100 million pair-wise associations. Break-up of the entity counts across the different entity types is provided in the Supplementary Table 1. The specialized COVID-19 subnetwork (corresponding to the COVID-19 related literature) consisted of ∼7.8 million pair-wise associations involving ∼43,000 entities.

3.1 Drugs Associated with COVID-19 from Literature

Well-known drug re-purposing candidates for COVID-19 in CoNetz include remdesivir, hydroxychloroquine and lopinavir-ritonavir combination, as well as convalescent plasma therapy to were seen to be strongly associated with COVID-19 disease node. Lesser known drug leads such as ciclesonide, baricitinib, fedratinib, tocilizumab and arbidol are also connected to the COVID-19 node. Alternative medicines mentioned in literature in the context of COVID-19 include astragali radix, shufeng jiedu and lianhuaqingwen. The top-15 chemicals and drugs directly associated with COVID-19 disease are shown in Figure 2. This shows the presence of high quality association neighbors in the network within the top-K neighborhood.

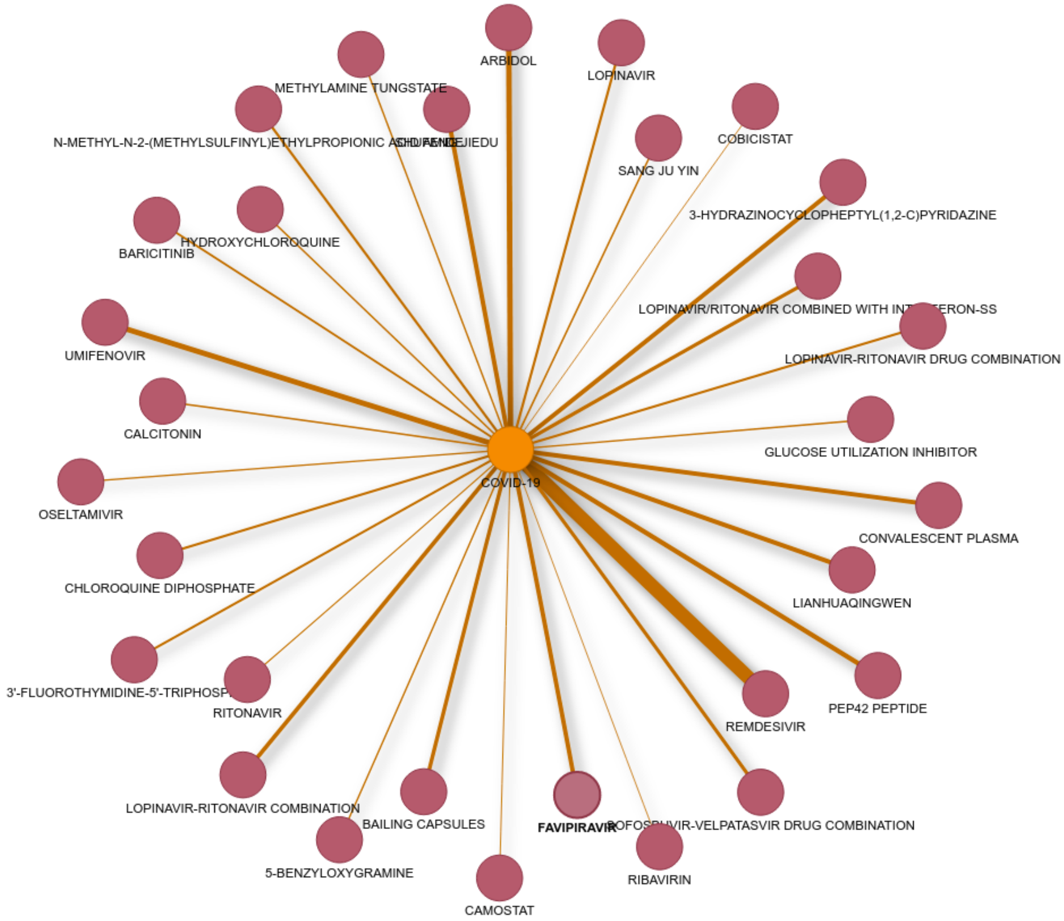
ARBIDOL
METHYLAMINE TUNGSTATE
LOPINAVIR
N-METHYL-N-2-(METHYLSULFINYL)ETHYLPROPIONIC ACID AMIDE
SHUFENG JIEDU
COBICISTAT
SANG JU YIN
3-HYDRAZINOCYCLOPHEPTYL(1,2-C)PYRIDAZINE
HYDROXYCHLOROQUINE
BARICITINIB
LOPINAVIR/RITONAVIR COMBINED WITH INTERFERON-SS
LOPINAVIR-RITONAVIR DRUG COMBINATION
UMIFENOVIR
CALCITONIN
GLUCOSE UTILIZATION INHIBITOR
COVID-19
OSELTAMIVIR
CONVALESCENT PLASMA
CHLOROQUINE DIPHOSPHATE
LIANHUAQINGWEN
3'-FLUOROTHYMIDINE-5'-TRIPHOSP
RITONAVIR
PEP42 PEPTIDE
REMDESIVIR
LOPINAVIR-RITONAVIR COMBINATION
BAILING CAPSULES
FAVIPIRAVIR
SOFOSBUVIR-VELPATASVIR DRUG COMBINATION
5-BENZYLOXYGRAMINE
RIBAVIRIN
CAMOSTAT

Figure 2: Top-15 chemicals and drugs directly associated with COVID-19 disease

3.2 Analysis of Novel Drug Predictions for COVID-19

The top-40 link-predicted drugs for COVID-19 in the network are shown in Figure 3. The list of these drugs along with additional details is provided in Supplementary Table 2.

We investigate these drugs as potential drug re-purposing candidates for COVID-19. Towards this, we perform network exploration around some of these link-predicted drugs to understand their possible mechanisms. We also identify and describe literature evidence that could support these predictions. While we provide analysis for a few of the 40 drugs, a similar analysis can be done for all 40 or any top-K drugs of ones choice. Our goal here is to illustrate the potential of the network and also enable domain experts to explore the network to generate their own hypotheses.

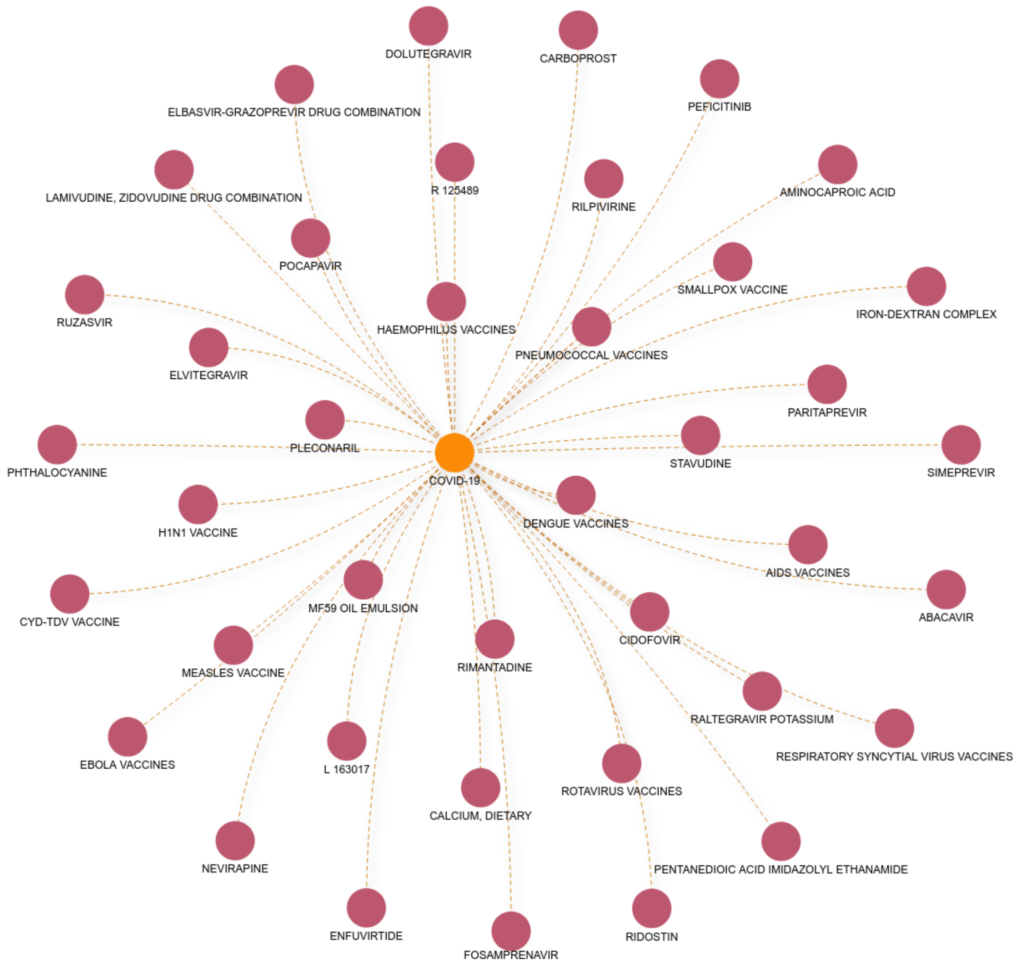

Figure 3: Top-40 chemicals and drugs predicted to be linked to COVID-19 disease. It is to be noted that these are not directly associated to COVID-19, but have been predicted to do so

1. **Ingavirin**: Ingavirin is a drug licensed in Russia for influenza and other acute respiratory viral infections [17, 18, 19]. It is a non-toxic broad spectrum antiviral with complex mechanism of action. In influenza, it is known to inhibit viral replication process by targeting the influenza nucleoprotein. In addition, experimental results suggest that it also interferes with the virus assembly process as well as budding leading to reduction of viral load. Experiments showed that increased viral clearance could be possibly due to the ability of the drug to overcome virus-induced immunosuppression [20].

As seen in Figure 4, ingavirin is strongly associated to the nucleoprotein (N Gene), which in turn is directly associated with the COVID-19 disease node and via the replicase polyprotein 1A node. The general physical properties of N proteins are markedly similar, and interestingly most of the characteristics of coronavirus N proteins are also shared by the nucleocapsid (NP) proteins of influenza viruses. Despite general similarities in the coronavirus N proteins, there exists only a low degree of sequence homology among them [21]. Based on this, ingavirin could be a potential COVID-19 drug, acting possibly against the SARS-CoV-2 nucleoprotein and inhibiting viral replication. There are reports in the general press that clinical trials are underway for testing the efficacy of ingavirin as a drug against general coronavirus infections [22].

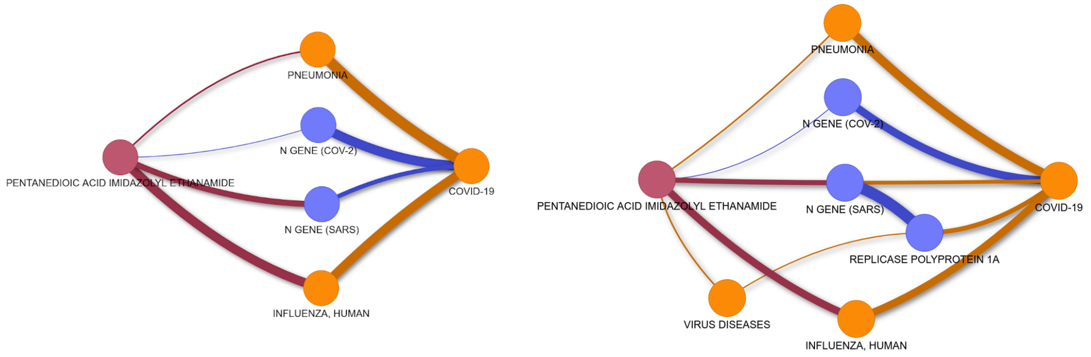

Figure 4: Network neighborhood for ingavirin, COVID-19 pair. The left side subnetwork shows the immediate neighborhood and the right side subnetwork shows the next level neighborhood. Ingavirin is associated with the COVID-19 disease node via the replicase polyprotein 1A node

2. **R 125489 (Laninamivir)**: Laninamivir is a neuraminidase inhibitor licensed for treatment of influenza A and B infections in different parts of the world [23]. Laninamivir has been prescribed as a single inhaled dose through its octanoate prodrug (CS-8958) against the viral neuraminidase [24, 25]. The neuraminidase enzyme seen in influenza A and B viruses cleaves the terminal sialic acid residues from carbohydrate moieties on the surfaces of host cells and virus envelope, resulting in the release of progeny viruses from infected cells [26, 27]. It is thought that laninamivir blocks the active site of the neuraminidase enzyme, leaving uncleaved sialic acid residues on the surfaces of host cells and influenza viral envelopes. This results in viral aggregation at the host cell surface when viral hemagglutinin binds to the uncleaved sialic acid residues, thus reducing the amount of virus that could be released to infect other cells [28].

   As seen in Figure 5, laninamivir is associated with sialic acid and signaling receptor binding function nodes, which in turn are associated with the COVID-19 disease node. It has been shown that the domain A of coronavirus S-proteins mediates attachment to oligosaccharide receptors such as 9-O-Ac-Sia depending on the coronavirus type [29]. The structural basis for human coronavirus attachment to sialic acid receptors at the surface of host cells has been established [30]. We conjecture that laninamivir can potentially affect the binding of SARS-CoV-2 to human host cells.

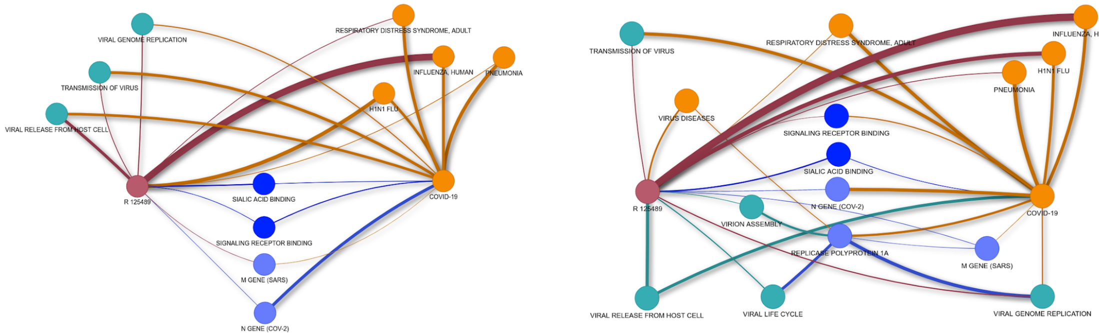

Figure 5: Network neighborhood for the laninamivir-COVID-19 pair. The left side subnetwork shows the immediate neighborhood and the right side subnetwork shows the next level neighborhood. Laninamivir is associated with sialic acid binding and signaling receptor binding function nodes, which in turn are associated with the COVID-19 node.

3. **Nevirapine**: Nevirapine is a non-nucleoside reverse transcriptase inhibitor used in the treatment of HIV [31]. Molecular docking studies have been carried out in order to search for potent SARS-CoV protease inhibitors [32]. This study showed nevirapine, glycovir, virazole and calanolide A to fit well at the substrate binding cleft of the 3D structure of the SARS-CoV protease, and that nevirapine with an estimated binding free energy of -9.47kcal/mol was a very good candidate [32]. This raises the possibility of nevirapine being a potential SARS-CoV-2 drug acting as SARS-CoV-2 main protease inhibitor. In fact, while this manuscript was under preparation, nevirapine has been proposed as a possible COVID-19 drug, based on it's potential action as a SARS-CoV-2 main protease inhibitor [6]. Though our corpus did not contain this article, CoNetz predicted a highly ranked link between nevirapine and COVID-19, thereby exhibiting the utility of our network-based approach. Figure 6 shows the network neighborhood between them.

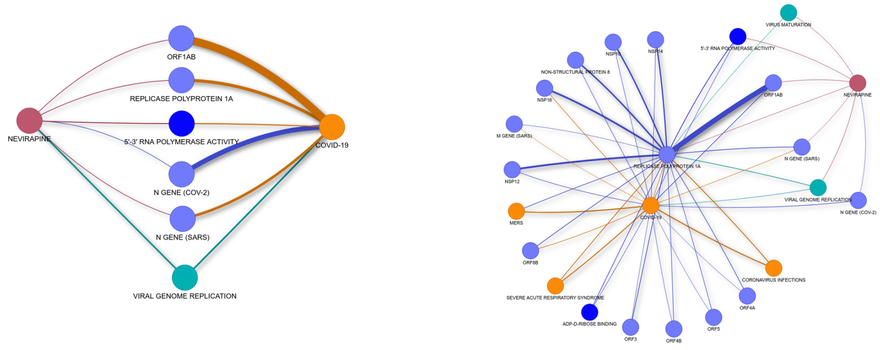

Figure 6: Immediate and next level network neighborhood for nevirapine-COVID-19 pair. Nevirapine could potential act as a SARS-CoV-2 main protease inhibitor.

4. **Paritaprevir**: Paritaprevir is an anti-hepatitis C direct-acting antiviral agent (DAA) that are viral NS3/NS4/NS4A protease inhibitors [33, 34, 35]. These non-structural proteins are vital for viral replication [34]. Treatment with DAA's such as Paritaprevir not only inhibits viral replication but may alter host adaptive and innate immune responses [34].

   As seen in Figure 7, paritaprevir is associated with nsp4 and viral genome replication. It has been shown that nsp4 is very crucial for general coronavirus-induced host membrane rearrangement and the replication complex function [36]. Furthermore, it was observed that the crucial amino acids residues (H120 and F121) in nsp4 are well conserved among betacoronaviruses, including MERS-CoV [36]. In fact, a very recent article not present in the corpus mentions Paritaprevir as a potential drug COVID-19 acting on the SARS-CoV-2 3C-like proteinase(3CLpro) [37], further showing the promise of link-prediction and network analysis.

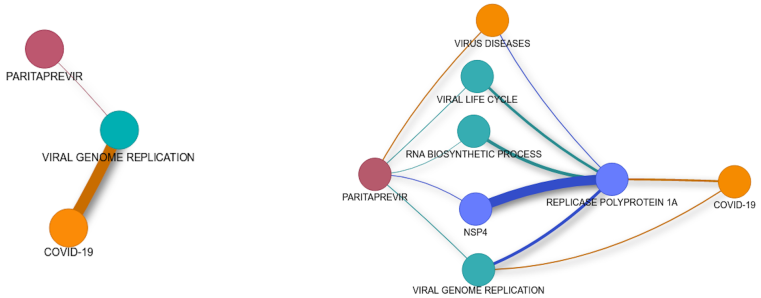

Figure 7: Immediate and next level network neighborhood for paritaprevir, COVID-19 pair. Paritaprevir is associated with nsp4 and viral genome replication.

We now investigate some of the non anti-viral drugs from Figure 3.

1. **Pranlukast**: Mannose-binding lectin (MBL) is an innate immune system protein that acts against a wide range of pathogenic microbes via complement activation [38, 39]. Clinical studies have shown that MBL deficiency might predispose one to severe respiratory tract infection [38]. The binding of MBL to pathogenic surfaces leads to the activation of associated MBL-associated serine proteases, production or cleavage of complement factors, amongst a host of downstream effects [38, 40]. This could potentially lead to potent bronchoconstriction and development of allergic inflammation.

   It has been previously reported that MBL contributes to the first-line host defense against SARS-CoV wherein deposition of complement C4 on SARS-CoV was enhanced by MBL [41]. Moreover, MBL deficiency has also been seen to be a susceptibility factor for acquisition of SARS [41]. Pranlukast is a leukotriene receptor antagonist and has an anti-inflammatory effect on bronchial eosinophilic infiltration [42]. In the context of coronaviruses, it has been proposed that a leukotriene C4 receptor antagonist such as pranlukast can be administered when a patient is diagnosed with SARS, thereby avoiding severe respiratory symptoms [43]. As seen in Figure 8, pranlukast is associated with COVID-19 disease via mannan binding and leukotriene C4, thus raising the possibility of pranlukast being used to manage the severe respiratory symptoms of COVID-19.

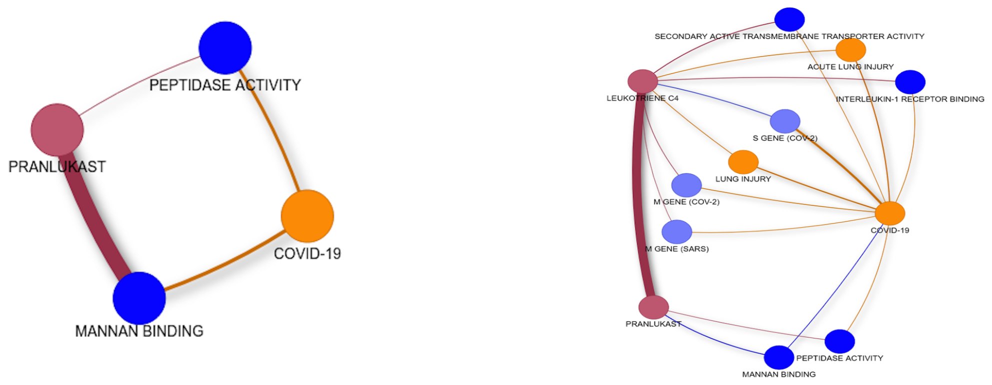

Figure 8: Immediate and next level network neighborhood for pranlukast, COVID-19 pair. Pranlukast is associated with COVID-19 disease via mannan binding and leukotriene C4.

2. **Peficitinib**: Peficitinib is an oral Janus kinase (JAK)1, JAK2, JAK3 and tyrosine kinase (Tyk)2 (pan-JAK) inhibitor approved in Japan for the treatment of rheumatoid arthritis [44]. Inhibition of JAK suppresses the activation of cytokine signaling pathways involved in inflammation and joint destruction in rheumatoid arthritis. Peficitinib suppresses the JAK-STAT pathway and the monocyte chemotaxis and proliferation of fibroblast-like synoviocytes through inhibition of inflammatory cytokines [45].

   As seen in Figure 9, peficitinib is linked to COVID-19 via baricitinib, and also via AAK1. Baricitinib is a numb-associated kinases (NAK) inhibitor that has a high affinity for AAK1, an important regulator of clathrin-mediated endocytosis [46]. The high affinity of baricitinib for NAKs, its anti-inflammatory properties, its ability to reduce associated chronic inflammation in interferonopathies and its advantageous pharmacokinetic properties appear to make it a special case among the drugs against COVID-19 [46]. Possibly, peficitinib might share some of these properties and hence might be effective in COVID-19 management. On the flip side, it is known that JAK inhibitors such as peficitnib can increase the risk of viral infections [47]. Hence, further investigation is needed to see if peficitinib could safely be used as a drug in COVID-19 treatment and management.

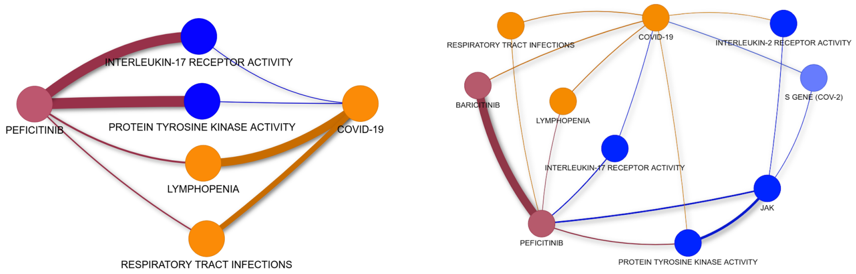
INTERLEUKIN-17 RECEPTOR ACTIVITY
PEFICITINIB
PROTEIN TYROSINE KINASE ACTIVITY
COVID-19
LYMPHOPENIA
RESPIRATORY TRACT INFECTIONS
RESPIRATORY TRACT INFECTIONS
COVID-19
INTERLEUKIN-2 RECEPTOR ACTIVITY
BARICITINIB
LYMPHOPENIA
S GENE (COV-2)
INTERLEUKIN-17 RECEPTOR ACTIVITY
JAK
PEFICITINIB
PROTEIN TYROSINE KINASE ACTIVITY

Figure 9: Immediate and next level network neighborhood for peficitinib, COVID-19 pair. Peficitinib is linked to COVID-19 via Baricitinib

4 Discussion

In this study, we tailored the TPX text-mining framework to extract pair-wise correlations from literature and built comprehensive correlation networks covering a set of 15 biomedical entity types. The entity types include human gene, SARS-CoV-2 gene, SARS gene, MERS gene, HPO phenotype, drug, disease symptom, chemical, disease, GO process, GO function, GO localization, cell type, tissue and anatomy. We also capture enriched pair-wise correlations in the context of COVID-19 and related articles and represent them as additional edge attributes in the network. Since the existing literature is an important source for effective link prediction, we applied GCAS link prediction algorithm on the network to predict novel associations. We illustrated the potential of the network by applying it to the problem of drug re-purposing for COVID-19 treatment and management. Nevirapine and paritaprevir, which were predicted in CoNetz to be linked to COVID-19, appeared in some very recent literature articles as potential COVID-19 drugs. However, these articles were not part of our corpus, which illustrates the utility of our network and approach.

To the best of our knowledge, CoNetz is the only network derived from literature that covers a wide range of 15 entity types. Existing publicly available networks are specific to a limited set of entities such as genes, proteins and diseases. Moreover, these networks lack up-to-date information. In contrast, CoNetz provides a rich neighborhood context for any subnetwork of interest by including a wide variety of entities. Furthermore, the network links have associated strength values that can be used to identify top-K strongly correlated neighbors of entities. We note that lack of comprehensive and curated high quality validation datasets makes quality evaluation of our approach difficult. We have previously performed quality evaluation for similar networks that were created in the context of rare-disease studies [11]. Analogous experiments involving validation datasets covering a wide range of entities would help in quality evaluation of the networks created in this study. It is to be noted that manual inspection of several top-K neighbors of COVID-19 related entities in the network showed significant fraction of highly relevant neighbors. CoNetz could be a valuable resource for COVID-19 studies carried out by the research community and the structured network data allows easy integration of the literature data to automated downstream analysis. In future, additional entities such as other related pathogen genes, as well as associations from high quality and heterogeneous curated data such as IntAct [48] with sufficient coverage can be explored for augmenting it.

CoNetz is compatible with the NetworkX and PyVis utilities. We provide several network exploration and visualization functionalities based on NetworkX and PyVis. We performed network exploration on the network to identify information related to COVID-19 disease. For exploration, a network node can be identified either by its “term name” or the entity ID in its source DB. The ID here is its public ID as present in the public data source from which its source dictionary was built. For instance, the Human Gene dictionary was built using HGNC. Hence, in order to query for “ACE2”, one could either search using its symbol “ACE2” or its HGNC ID “HGNC:13557”.

With the sudden and quick spread of the COVID-19 across the world, there is a great need to understand the pathobiology of the SARS-CoV-2 virus and develop drugs to treat and manage COVID-19 disease. While development of new drugs would take time, repurposing of already available drugs could be an faster alternative. Automated extraction of useful information from literature in an up-to-date manner and automated generation of novel insights is crucial for accelerated discovery of drug/vaccine targets and re-purposing drug candidates. Our literature driven text and network mining based approach constructs an up-to-date knowledge base from literature that we update periodically. CoNetz provides a promising collection of COVID-19 drug re-purposing candidates. While we provide supporting arguments for these predicted drugs, these have to be further investigated using other methods such as expression studies and also experimental studies to identify high confidence drug candidates.

5 Conclusions

Our automated text and network-mining approach builds an up-to-date and comprehensive knowledge network CoNetz from literature for COVID-19 studies. Furthermore, the wide range of entity types captured in CoNetz provides a rich neighborhood context for any subnetwork of interest. Our approach avoids multiple drawbacks associated with manual or semi-manual curation including cost and effort involved, lack of up-to-date information and limited coverage. CoNetz can be a valuable resource for COVID-19 studies carried out by the research community. Furthermore, the structured network data allows easy integration of the literature data to automated downstream analysis. As an example, the network predicted several promising and novel repurposing candidates for COVID-19 drugs, including ingavirin, laninamivir, nevirapine, paritaprevir, pranlukast and peficitinib. Literature investigation of these candidates reveal that they can target different mechanisms and may merit investigation for COVID-19 treatment.

List Of Abbreviations

Abbreviation	Full Form
COVID-19	coronavirus disease 2019
SARS	severe acute respiratory syndrome
MERS	Middle East respiratory syndrome
SARS-CoV	severe acute respiratory syndrome coronavirus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
HANRD	Heterogeneous Association Network for Rare Diseases
GCAS	Graph Convolution-based Association Scoring
NER	Named Entity Recognition
TPX	TCS Pubmed eXplorer
PPI	Protein-Protein Interactions
CR	Conflict Resolver

Declarations

5.1 Ethics approval and consent to participate

Not applicable

5.2 Consent for publication

Not applicable

Availability of data and material

The authors declare that all data supporting the findings of this study are available within the article and its supplementary file. We have made available CoNetz and utilities at https://web.rniapps.net/tcn/tcn.tar.gz

Competing interests

The authors declare that they have no competing interests.

Funding

All the authors are permanent employees of Tata Consultancy Services Ltd.

Authors' Contributions

AR worked on the overall ideas, performed analysis of input data sources and drafted the manuscript. SV worked on algorithm design, implemented various algorithms, performed analysis of data, created the visualization, and gave inputs for the manuscript. TJ contributed to the biomedical content, biological analysis of the output and drafted the manuscript. SK helped in creating the networks, performing data analysis and packaging the entire study code. NS worked on analyzing various algorithms, helped develop thresholds for the parameters and drafted the manuscript. RS oversaw the development of the entire study, reviewed all implementations that were used and gave crucial biological insights. All authors have read and approved the final manuscript.

References

[1] Senanayake SL. Drug repurposing strategies for COVID-19. Future Drug Discov. 2020;.

[2] Lake MA. What we know so far: COVID-19 current clinical knowledge and research. Clinical Medicine. 2020;20(2):124–127.

[3] Zhou Y, Hou Y, Shen J, Huang Y, Martin M, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discovery. 2020;6:14.

[4] Gordon DE, et al. A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. bioRxiv. 2020;2020.03.22.002386.

[5] Gysi DM, et al. Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19. arXiv. 2020;2004.07229.

[6] Wang J. Fast Identification of Possible Drug Treatment of Coronavirus Disease -19 (COVID-19) Through Computational Drug Repurposing Study. J Chem Inf Model. 2020;.

[7] Zeng X, Zhu S, Hou Y, Zhang P, Li L, Li J, et al. Network-based Prediction of Drug-Target Interactions using an Arbitrary-Order Proximity Embedded Deep Forest. Bioinformatics. 2020;.

[8] CORD-19 Dataset: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge;.

[9] Joseph T, Saipradeep VG, Raghavan GSV, Srinivasan R, Rao A, Kotte S, et al. TPX: Biomedical literature search made easy. Bioinformation. 2012;8(12):578.

[10] Rao A, Saipradeep VG, Joseph T, Kotte S, Sivadasan N, Srinivasan R. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. BMC Medical Genomics. 2018;11(1):57.

[11] Rao A, Joseph T, Saipradeep VG, Kotte S, Sivadasan N, Srinivasan R. PRIORI-T: A tool for rare disease gene prioritization using MEDLINE. PLoS One. 2020;15(4):e0231728.

[12] Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88:265.

[13] Gene Ontology Consortium. Gene ontology consortium: going forward. Nucleic Acids Research. 2015;43(Database issue):D1049–D1056.

[14] Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames. org: the HGNC resources in 2015. Nucleic Acids Research. 2015;43(Database issue):D1079.

[15] NetworkX: https://pypi.org/project/networkx/;.

[16] VisJS: https://almende.github.io/vis/;.

[17] Semenova NP, Prokudina EN, Livov DK, Nebol'sin VE. Effect of the antiviral drug Ingavirin on intracellular transformations and import into the nucleus of influenza A virus nucleocapsid protein. Vopr Virusol. 2010;55:17–20.

[18] Zarubaev VV, Belyaevskaya SV, Sirotkin AC, Anfimov PM, Nebol'sin VE, Kiselev OI, et al. Effect of Ingavirin on ultrastructure and infectivity of influenza virus in vitro and in vivo. Vopr Virusol. 2011;56:21–25.

[19] Zarubaev VV, Garshinina AV, Kalinina NA, Shtro AA, Belyaevskaya SV, Slita AV, et al. Activity of Ingavirin (6-[2-(1H-Imidazol-4-yl)ethylamino]-5-oxo-hexanoic Acid) Against Human Respiratory Viruses in in Vivo Experiments. Pharmaceuticals (Basel). 2011;4:1518–1534.

[20] Kuznetsova I, Egorov A, Aschacher T, Nibolsin V, Bergmann M. Novel antiviral drug ingavirin restores the cellular antiviral response in influenza A virus infection and enhances viral clearance in ferrets. In: 3rd International Influenza Meeting; 2012. p. 170.

[21] Laude H, Masters P. The Coronavirus Nucleocapsid Protein; 1995.

[22] https://en.currenttime.tv/a/infodemic-of-fake-coronavirus-info-spreads-throughout-russia-ukraine-central-asia/30413076.html;.

[23] Gubareva L, Mohan T. Antivirals Targeting the Neuraminidase. Cold Spring Harb Perspect Med. 2020;pii:a038455.

[24] Vavricka CJ, Li Q, Wu Y, Qi J, Wang M, Liu Y, et al. Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for influenza NA inhibition. PLoS Pathogy. 2011;7(10):e1002249.

[25] Kubo S, Kakuta M, Yamashita M. In vitro and in vivo effects of a long-acting anti-influenza agent CS-8958 (laninamivir octanoate, Inavir) against pandemic (H1N1) 2009 influenza viruses. Jpn J Antibiot. 2010;63(5):337–346.

[26] Calfee DP, Hayden F. New approaches to influenza chemotherapy: neuraminidase inhibitors. Drugs. 1998;56:537–553.

[27] Liu C, Eichelberger MC, Compans RW, M AG. Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly or budding. J Virol. 1995;69:1099–1106.

[28] Palese P, Compans RW. Inhibition of influenza virus replication in tissue culture by 2-deoxy-2,3-dehydro-N-trifluoroacetylneuraminic acid (FANA): mechanism of action. J Gen Virol. 1976;33:159–163.

[29] Li W, Hulswit RJG, Widjaja I, Raj VS, McBride R, Peng W, et al. Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. Proc Natl Acad Sci. 2017;114(40):E8508–E8517.

[30] Tortorici MA, Walls AC, Lang Y, Wang C, Li Z, Koerhuis D, et al. Structural basis for human coronavirus attachment to sialic acid receptors. Nat Struct Mol Biol. 2019;26(6):481–489.

[31] Tateishi Y, Ohe T, Yasuda D, Takahashi K, Nakamura S, Kazuki Y, et al. Synthesis and evaluation of nevirapine analogs to study the metabolic activation of nevirapine. Drug Metab Pharmacokinet. 2020;35(2):238–243.

[32] Lee VS, Wittayanarakul K, Remsungnen T, Parasuk V, Sompornpisut P, Chantratita WC, et al. Structure and Dynamics of SARS Coronavirus Proteinase: The Primary Key to the Designing and Screening for Anti-SARS Drugs. Science Asia. 2010;p. 181–188.

[33] McConachie SM, Wilhelm SM, Kale-Pradhan PB. New direct-acting antivirals in hepatitis C therapy: a review of sofosbuvir, ledipasvir, daclatasvir, simeprevir, paritaprevir, ombitasvir and dasabuvir. Expert Rev Clin Pharmacol. 2016;9(2):287–302.

[34] Geddawy A, Ibrahim YF, Elbahie NM, Ibrahim MA. Direct Acting Anti-hepatitis C Virus Drugs: Clinical Pharmacology and Future Direction. J Transl Int Med. 2017;5(1):8–17.

[35] Loo N, Lawitz E, Alkhouri N, Wells J, Landaverde C, Coste A, et al. Ombitasvir/paritaprevir/ritonavir + dasabuvir +/- ribavirin in real world hepatitis C patients. World J Gastroenterol. 2019;25(18):2229–2239.

[36] Sakai Y, Kawachi K, Terada Y, Omori H, Matsuura Y, Kamitani W. Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication. Virology. 2017;510:165–174.

[37] Khan RJ, Jha RK, Amera GM, Jain M, Singh E, Pathak A, et al. Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase. J Biomol Struct Dyn. 2020;20:1–14.

[38] Eisen DP. Mannose-binding lectin deficiency and respiratory tract infection. J Innate Immun. 2010;2(2):114–122.

[39] Turner MW. Mannose-binding lectin (MBL) in health and disease. Immunobiology. 1998;199(2):327–339.

[40] Dunkelberger JR, Song WC. Complement and its role in innate and adaptive immune responses. Cell Res. 2010;20(1):34–50.

[41] Ip WK, Chan KH, Law HK, Tso GH, Kong EK, Wong WH, et al. Mannose-binding lectin in severe acute respiratory syndrome coronavirus infection. J Infect Disease. 2005;191:1697–1704.

[42] Yoshida S, Ishizaki Y, Shoji T, Onuma K, Nakagawa H, Nakabayashi M, et al. Effect of pranlukast on bronchial inflammation in patients with asthma. Clin Exp Allergy. 2000;30(7):1008–1014.

[43] Nozaki M. Method of treating or inhibiting the development of brain inflammation and sepsis; 2006. US Patent US7148248B2.

[44] Markham A, Keam SJ. Peficitinib: First Global Approval. Drugs. 2019;79(8):887–891.

[45] Ikari Y, Isozaki T, Tsubokura Y, Kasama T. Peficitinib Inhibits the Chemotactic Activity of Monocytes via Proinflammatory Cytokine Production in Rheumatoid Arthritis Fibroblast-Like Synoviocytes. Cells. 2019;8(6):E561.

[46] Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, et al. COVID-19: combining antiviral and anti-inflammatory treatments. The Lancet Infectious Diseases. 2020;20(4):400–402.

[47] You H, Xu D, Zhao J, Li J, Wang Q, Tian X, et al. JAK Inhibitors: Prospects in Connective Tissue Diseases. Clin Rev Allergy Immunol. 2020;.

[48] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research. 2013;42(Database issue):D358–D363.