Learning COVID-19 Network from Literature Databases Using Core Decomposition

by

Yang Guo
B.Eng., Huazhong Agricultural University, 2019

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Learning COVID-19 Network from Literature Databases Using Core Decomposition

by

Yang Guo
B.Eng., Huazhong Agricultural University, 2019

Supervisory Committee

Dr. Xuekui Zhang, Co-Supervisor
(Department of Mathematics and Statistics)

Dr. Li Xing, Co-Supervisor
(Department of Mathematics and Statistics)

Dr. Venkatesh Srinivasan, Outside Member
(Department of Computer Science)

ABSTRACT

The SARS-CoV-2 coronavirus is responsible for millions of deaths around the world. To help contribute to the understanding of crucial knowledge and to further generate new hypotheses relevant to SARS-CoV-2 and human protein interactions, we make use of the information abundant Biomine probabilistic database and extend the experimentally identified SARS-CoV-2-human protein-protein interaction (PPI) network *in silico*. We generate an extended network by integrating information from the Biomine database and the PPI network. To generate novel hypotheses, we focus on the high-connectivity sub-communities that overlap most with the PPI network in the extended network. Therefore, we propose a new data analysis pipeline that can efficiently compute core decomposition on the extended network and identify dense subgraphs. We then evaluate the identified dense subgraph and the generated hypotheses in three contexts: literature validation for uncovered virus targeting genes and proteins, gene function enrichment analysis on subgraphs, and literature support on drug repurposing for identified tissues and diseases related to COVID-19. The majority types of the generated hypotheses are proteins with their encoding genes and we rank them by sorting their connections to known PPI network nodes. In addition, we compile a comprehensive list of novel genes, and proteins potentially related to COVID-19, as well as novel diseases which might be comorbidities. Together with the generated hypotheses, our results provide novel knowledge relevant to COVID-19 for further validation.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My parents, for supporting me in the low moments.

Professors Xuekui Zhang and Li Xing, for mentoring, support, encouragement, and patience.

Professor Venkatesh Srinivasan, for his dedicated help and valuable suggestions on my research and for agreeing to serve on my thesis committee.

Professor Lin Cai, for agreeing to be my thesis external examiner and for her constructive comments and recommendations.

The Visual and Automated Disease Analytics (VADA) program, for funding me with a fellowship.

All the staff in the department, for the countless help they provided.

And many strokes, though with a little axe / Hew down and fell the hardest-timbered oak.

William Shakespeare,"King Henry VI Part III", Act 2 scene 1

Greatest English dramatist & poet (1564 - 1616)

Chapter 1

Introduction

The COVID-19 pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a grave threat to public health and the global economy. It has led to more than 178 million confirmed cases and 3.85 million deaths worldwide as of June 18, 2021. SARS-CoV-2 is a newly discovered positive-sense single-stranded RNA virus and belongs to the member of the Coronaviridae (CoV) family [88]. It shares 89.1% and 50% nucleotide similarity to other previously detected human coronavirus SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-Cov) [21, 76, 54], respectively. Previous studies have also shown that they share stronger similarities with respect to their structures and pathogenicity [54]. These provide valuable knowledge to facilitate our understanding of the pathophysiology of SARS-CoV-2. Particularly, we now know that SARS-CoV-2 mainly enters human cells via binding of its spike protein to the angiotensin converting enzyme 2 (ACE2) receptor [35] and is associated with an extensive immune reaction referred to as "cytokine storm" triggered by the excessive production of interleukin 1 beta (IL-1b), interleukin 6 (IL-6), and others. However, much remains to be explored about how these critical human proteins are involved in the infection and the associated COVID-19 pathology [21], critical towards devising therapeutic strategies to counteract SARS-CoV-2 infection.

In order to investigate the complications and comorbidities of SARS-CoV-2 and to facilitate the search for effective treatment, many studies have been conducted to investigate the host dependencies of the SARS-CoV-2 virus from a systems level. For example, Blanco-Melo performed a comparative transcriptional analysis of COVID-19 patients responding to SARS-CoV-2 and other respiratory viruses, that revealed reduced innate antiviral defenses coupled with exuberant inflammatory cytokine pro-

duction as the defining and driving features of COVID-19 [10]. Bojkova et al. conducted proteomic analysis to identify the host cell pathways that are modulated by SARS-CoV-2 and showed that inhibition of these pathways prevents viral replication in human cells [11]. Gordon et al. systematically mapped the interaction landscape between SARS-CoV-2 proteins and human proteins using affinity-purification mass spectrometry [34]. They identified 332 high-confidence protein interactions between SARS-CoV-2 viral proteins and human proteins related to various complexes and biological processes (about 40% of human proteins identified to interact with SARS-CoV-2 were associated with endomembrane system or membrane vesicle trafficking). From the presented SARS-CoV-2-human protein-protein interaction (PPI) network the authors identified 62 druggable SARS-CoV-2-interacting human proteins with 69 targeting ligands (drugs). All these studies contributed to a better understanding of the SARS-CoV-2 and host protein interactome, providing insights for the development of therapies for the treatment of COVID-19. However, these studies only revealed different aspects of the potential mechanisms behind SARS-CoV-2 infection at specific conditions and do not bring out into open the comorbidities-, cell- and organ-type-specific human-viral interaction architecture.

To generate new hypotheses, we are interested in extending the SARS-CoV-2 viral-human protein interaction (PPI) network discovered by Gordon et al. [34] through the means of integrative network analysis from publicly available research databases. We aim to identify high-connectivity sub-communities (connected with PPI network) from the integrated dataset since our fundamental assumption is that members in such sub-communities play more important roles in the network [44, 71].

1.1 Approaches Using Integrative Network Analysis

Integrative network analysis provides an efficient approach to enable discovery and evaluation of (unknown) connections spanning multiple types of relationships inferred from different omics studies. Recently, Zhou et al. [86] and Kumar et al. [44] have applied similar ideas to perform the integrative network analysis for elucidating the molecular mechanisms of SARS-CoV-2 pathogenesis. Different from their studies based on limited experimental datasets, here we adopt a large probabilistic biological network called Biomine [31, 53] which integrates several databases including PubMed

[72], UniProt [25], STRING [38], Entrez Gene [47, 72], and InterPro [37]. Many biological networks can be represented using probabilistic graph structures due to the intrinsic uncertainty present in their measurements. For instance, the edges in PPI networks obtained through laboratory experiments are often prone to measurement errors. The edges are often labeled with uncertainty levels that can be interpreted as probabilities. We aim to mine these probabilistic graphs to enhance our understanding of the SARS-CoV-2 and human protein interactions and to further aid the discovery of the essential/unknown knowledge relevant to the interactions between hosts and SARS-CoV-2 virus. The crux of our approach is to use a *core decomposition* strategy that detects highly connected sub-communities. Unlike other notions of cohesive subgraphs, e.g. *cliques*, *n-cliques*, *k-plexes*, which are all **NP**-hard to compute, *k-core* can be computed in polynomial time [13, 43, 32]. The goal of *k-core* computation is to identify the largest induced subgraph of a graph G in which each vertex is connected to at least k other vertices. The set of all *k-core*s of G forms the core decomposition of G [62]. The coreness (core number) of a vertex v in G is defined as the maximum k such that there is a *k-core* of G containing v. For an example of core decomposition, see Figure 1.1. For probabilistic graphs, the notion of core decomposition evolves into the more challenging probabilistic core decomposition. In this study, we connect the Biomine network to the small PPI network and we integrate our previously proposed graph peeling algorithm [32] for probabilistic core decomposition and proposed an analysis pipeline to detect the probabilistic coreness in data, finding the high-connectivity sub-communities, and generate hypotheses on COVID-19 relevant bio-networks.

Specifically, from the results, we are particularly interested in dense cores in the Biomine database that overlap with the PPI network as much as possible. The dense cores in a different region of the Biomine network that do not overlap with the PPI network are not of interest to us.

1.2 Our Contributions

1. **Pipeline on *k-core* decomposition and bio-network analysis**. A two-stage data screening procedure was added to the peeling algorithm (PA), making it focus more on cores with higher density. Overall, the pipeline consists of three steps: data preprocessing, PA, and functional enrichment analysis on the generated networks.

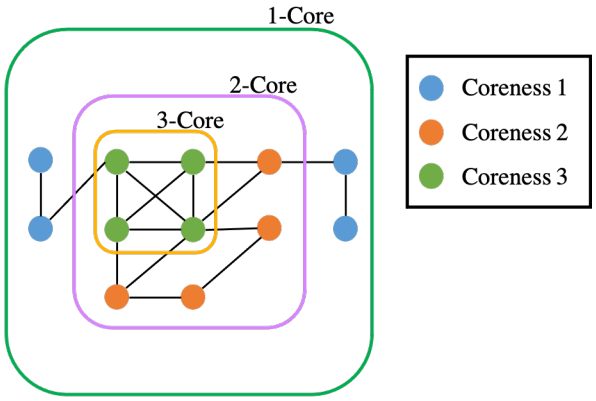


Figure 1.1: 3-core decomposition for an example graph

2. **Evaluation through literature mining and gene set enrichment analysis**. We evaluated the extended COVID-19 biological network in three contexts: literature support for identified tissues and diseases related to COVID-19, literature validation for uncovered SARS-CoV-2 targeting genes and proteins, and gene ontology (GO) over-representation test on the selected network for biological processes linked with RNA processing and viral transcription.

3. **COVID-19 hypotheses generation**. We discovered novel diseases that might be comorbidities, genes, and proteins that could potentially relate to COVID-19 and we presented them as top candidates for future validation.

Chapter 2

Methods and Materials

2.1 Biomine Database and SARS-CoV-2-Host Protein-Protein Interaction Network

The Biomine database is a large probabilistic biological network constructed using selected publicly available databases, for example, Entrez Gene, UniProt, STRING, InterPro, PubMed, Gene Ontology (GO), etc. The full Biomine database has 1508587 nodes, 32761889 edges and contains biology information of several species including humans. The SARS-CoV-2-host protein-protein interaction (PPI) network, identified by Gordon et al. [34], contains 692 nodes (27 SARS-CoV-2 viral proteins and 665 viral-interacting human proteins) and 695 edges. Since we are working with Homo sapiens data, as a preliminary stage of data screening, we select a subset of the full Biomine database, the *human* organism as the database to be used to extend the PPI network. This will eliminate approximately 43% of the full Biomine database. The *human* Biomine database contains 861812 nodes, 8666287 edges, and each entry possesses the form $\mathcal{E} = (from, to, relationship, link\_goodness)$. Here, $from$ and to are two nodes forming an edge in the network, $relationship$ is the link type describing the relationship between the two nodes, $link\_goodness$ is computed based on relevance, informativeness, and reliability [31] and is interpreted as the probability that the edge exists.

We then extend the PPI network with the *human* Biomine database and deal with duplicated entries and loops. For a detailed overview of the extension of SARS-CoV-2 and host protein-protein interaction network as well as duplicates/loops removal illustration, see next section.

2.2 SARS-CoV-2 and Host Protein-Protein Interaction Network Extension

In this section we describe the procedure of extending the SARS-CoV-2-host protein-protein interaction (PPI) network identified by Gordon et al. [34] with the *human* Biomine database. And we also introduce the pre-processing steps for duplicated entries and loops in the extended dataset.

Each entry in the *human* Biomine database possesses the following form:

$$\mathcal{E} = (from, to, relationship, link\_goodness)$$

For edges in the PPI network, we set *relationship* to be *COVID_before_stage1*, meaning they are predefined SARS-CoV-2 protein interactions; *link_goodness* is set to be 1.0 representing the edge as known with full confidence. For the *human* organism set, if two entries have the same $from$ and to nodes but with different *relationship* and *link_goodness*, we regard them as (semi-)duplicates and merge them into a single entry with the second *relationship* appended after the first and the new *link_goodness* being the average of the two previous *link_goodness* value.

We now append the formatted PPI network after the *human* Biomine data. This operation will surely introduce new duplicated entries in the merged dataset. We will re-do the duplicate removal operation on the new dataset. However, when merging the newly found duplicates, the new *link_goodness* will be the maximum *link_goodness* between the two to account for the 100% confidence edges in the PPI network. After duplicate removal, an additional loop removal procedure will be applied. Currently, our peeling algorithm (PA) can only operate on undirected graphs, hence any two edges with opposite direction (forming a loop) will be regarded as the same and processed. The maximum *link_goodness* between the two edges will be assigned to the processed edge. The idea of duplicate removal and loop removal is illustrated in Figure 2.1.

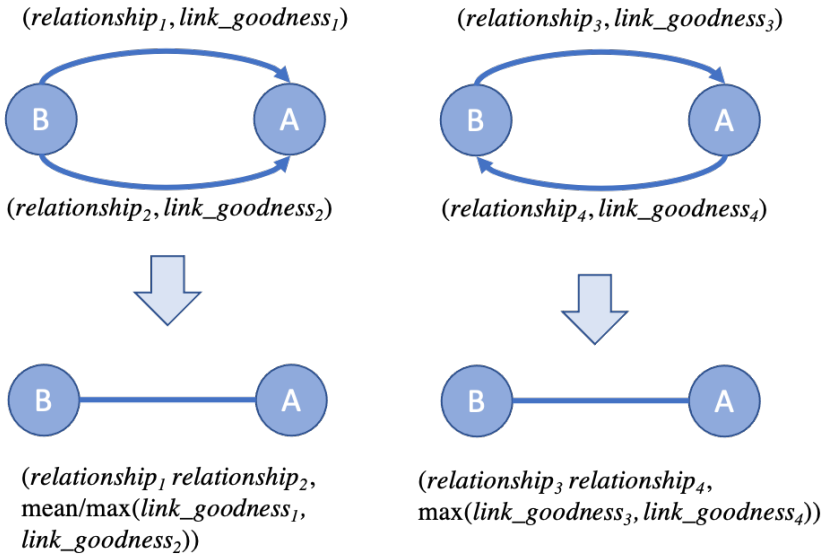


Figure 2.1: Remove duplicated and looped edges

2.3 Definition of Deterministic and Probabilistic Core Decomposition

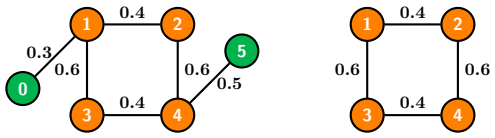
Let $G = (V, E)$ be an undirected graph, where V and E are the set of vertices and edges in G, respectively. Given a vertex $u \in V$, let $N_G(u)$ be the set of all neighbors of u, i.e. $N_G(u) = \{v : (u, v) \in E\}$. $|N_G(u)|$, is equal to deterministic degree of u in G.

Core decomposition in deterministic graphs. Given a graph G, the k-core of G is defined as the largest subgraph $H \subseteq G$ in which each vertex has degree of at least k in H. The set of all k-cores forms the core decomposition of G, where $0 \leq k \leq d_{\max}(G)$, and $d_{\max}(G)$ is the maximum vertex degree in G. Given a vertex u, the largest value of k for which u belongs to a k-core is called core number of u.

Probabilistic graphs. A probabilistic graph $\mathcal{G} = (V, E, p)$, is defined over a set of vertices V, a set of edges E and a probability function $p : E \rightarrow (0, 1]$ which assigns an existence probability $p(e)$ to every edge $e \in E$. For each $v \in V$, the number of edges incident on v in a deterministic graph is denoted by d_v and referred to as the deterministic degree of v. In the literature, the existence probability of each edge is assumed to be independent of other edges [13].

The *possible worlds* of $\mathcal{G}$ are deterministic graph instances of $\mathcal{G}$, which are used for

analyzing probabilistic graphs. In each possible world only a subset of edges appears. For each possible world $G = (V, E_G) \sqsubseteq \mathcal{G}$, where $E_G \subseteq E$, the probability of observing that possible world is obtained as follows: $\Pr(G) = \prod_{e \in E_G} p(e) \prod_{e \in E \setminus E_G} (1 - p(e))$.



(a) (b)

Figure 2.2: a) Probabilistic graph $\mathcal{G}$, b) (2,0.2)-core $\mathcal{H}$ of $\mathcal{G}$.

Let u be a vertex in $\mathcal{G}$. The probability that u has degree at least t in $\mathcal{G}$ can be expressed as $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq t] = \sum_{G \sqsubseteq \mathcal{G}} \Pr(G) \cdot \mathbb{1}(G, u, t)$, where $\mathbb{1}(G, u, t)$ is an indicator function which takes on 1 if degree of u in possible world G is at least t. It should be noted that as t decreases (increases), $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq t]$ increases (decreases). Given a user-defined threshed $\eta \in [0, 1]$, the η-degree of u [13], denoted by η-$\mathsf{deg}_{\mathcal{G}}(u)$, is defined as the maximum integer $t \in [0, d_u]$ for which $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq t] \geq \eta$, where d_u is the number of all edges incident to u which is equal to the deterministic degree of u.

Core decomposition in probabilistic graphs. We use the notion of (k, η)-core in [13] for core decomposition in probabilistic graphs. Let $\mathcal{G} = (V, E, p)$ be a probabilistic graph, and $\eta \in [0, 1]$ be a user-specified threshold. The (k, η)-*core* is the largest subgraph $\mathcal{H}$ of $\mathcal{G}$ in which each vertex u has η-degree no less than k, i.e. η-$\mathsf{deg}_{\mathcal{H}}(u) \geq k$.

Core decomposition of $\mathcal{G}$ is the set of all (k, η)-cores, for $k \in [0, k_{\max,\eta}]$, where $k_{\max,\eta} = \max_u\{\eta\text{-}\mathsf{deg}_{\mathcal{G}}(u)\}$. The *core number* of a vertex u, $\kappa_\eta(u)$, is the largest integer k for which u belongs to a (k, η)-core.

The set of all (k, η)-cores is the unique core decomposition of $\mathcal{G}$ and follows the following relation [32]:

$$\mathcal{G} = \mathcal{G}'_{(0,\eta)} \subseteq \mathcal{G}'_{(1,\eta)} \subseteq ... \subseteq \mathcal{G}'_{(k_{\max}-1,\eta)} \subseteq \mathcal{G}'_{(k_{\max},\eta)} \tag{2.1}$$

where $k_{\max}$ is the maximum probabilistic coreness of any vertex in $\mathcal{G}$.

For simplicity, we will use core number and coreness instead of η-core number/probabilistic coreness in the rest of the paper. Additionally, we use *dense* to describe cores with high core numbers and we use *large* to represent cores that have many nodes.

Consider Figure 2.2a, vertex $u = 1$, and $\eta = 0.2$. We have $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq 3] = 0.3 \cdot 0.4 \cdot 0.6 = 0.072$ (product of probabilities that edges $(0, 1)$, $(1, 2)$, and $(1, 3)$ exist), and $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq 2] = 0.396$. Since 0.396 is greater than η, $\eta\text{-}\mathsf{deg}_{\mathcal{G}}(u) = 2$.

Figure 2.2b shows a $(2, 0.2)$-core $\mathcal{H}$ of $\mathcal{G}$. Each vertex $u \in \mathcal{H}$, has η-degree 2 with probability 0.2.

Consider $u = 1$ and $\eta = 0.2$. Vertex u is in $(1, 0.2)$-core ($\mathcal{G}$ itself) and $(2, 0.2)$-core ($\mathcal{H}$). There is no $(3, 0.2)$-core, thus, $\kappa_\eta(u) = 2$.

η-degree computation using dynamic programming (DP). To find η-degree of each vertex u, $\Pr[\mathsf{deg}_{\mathcal{G}}(u) \geq t]$ should be computed. These probabilities can be computed using dynamic programming as proposed in [13].

2.4 Data Analysis Pipeline

The Biomine database contains abundant biological information. Though we used the smaller *human* Biomine database, after extending it with the SARS-CoV-2 protein-protein interaction (PPI) network the resulting network is still enormous and hard to reason. To make sense of such a huge network, we propose a data analysis pipeline with three steps described below.

2.4.1 Step 1: Data preprocessing

The data preprocessing step contains three sub-steps, screening based on degree expectation, screening based on lower-bounds of *η-degree*, and nodes retaining. The goal of this data preprocessing step is to reduce the nodes in the network and speed up the follow-up analyses. Specifically, we remove large proportion of nodes with too low connectivity to be members of dense sub-communities in the network.

Here we start with the first sub-step and we briefly explain the methodology and the rationale behind it. For each vertex $v \in V$, we have a set of edges incident to v and each edge is accompanied with a probability of existence p_i that is independent of other edge probabilities in $\mathcal{G}$. Given a probabilistic graph $\mathcal{G}$, and a vertex v, $deg_{\mathcal{G}}(v)$ can be interpreted as the sum of a set of independent Bernoulli random variables X_i's with different success probabilities p_i's [32] where:

$$X_i = \begin{cases} 1, & \text{if edge } e_i \text{ incident to } v \text{ exists in the graph} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

and $deg_{\mathcal{G}}(v)$ follows Poisson binomial distribution with $E[deg_{\mathcal{G}}(v)] = \sum E[X_i] = \sum p_i$. $\sum p_i$ therefore can be seen as a upper-bound to $deg_{\mathcal{G}}(v)$ so we use $\sum p_i$ as the first screening criteria. As thresholds are user-defined, any positive integer greater or equal to 0 is accepted but it is recommended that the first threshold is larger than 5 (if the first threshold is set to 5 it means any nodes with an expectation of degree less than 5 is removed, e.g. only nodes with $\sum p_i \geq 5$ and $\sum(1-p_i) \geq 5$ are kept). The goal of this step is to remove nodes that are rarely connected with others and hence are not eligible to be part of any highly connected sub-network. For example, if a vertex u has $\sum p_i$ less than 5, it means that the upper bound of $deg_{\mathcal{G}}(u)$ is less than 5, and hence u will not appear in cores with high activities. If the upper-bound $\sum p_i$ is lower, more nodes are retained. To speed up subsequent analyses, the threshold value should be high enough, yet it cannot be so high that possible highly connected nodes are removed. In our experiment, we empirically choose a conservative number, 5, as the first threshold, but other threshold values could be used.

For our merged dataset, 28224 nodes passed the screening.

Now, we introduce the second step of data prepossessing. For the 28224 nodes that passed the first stage of data screening, we calculate lower-bounds of their η-*degree* using Lyapunov Central Limit Theorem (CLT) implemented in PA. Given a vertex $v \in V$, based on Lyapunov CLT, $Z = \frac{1}{\sigma}\sum_{i=1}^{d_v}(X_i - \mu_i)$ has standard normal distribution, where $\mu_i = \Pr(X_i)$, and $\sigma = \sqrt{\sum_{i=1}^{d_v}\Pr(X_i)(1-\Pr(X_i))}$. Approximation of $\Pr[\mathsf{deg}_{\mathcal{G}}(v) \geq t] = \Pr[\sum_{i=1}^{d_v} X_i \geq t]$ can be obtained by subtracting μ_i from the sum of X_i's, and dividing by σ. As a result, we have:

$$\Pr\left[\sum_{i=1}^{d_v} X_i \geq t\right] = \Pr\left[\frac{1}{\sigma}\sum_{i=1}^{d_v}(X_i - \mu_i) \geq \frac{1}{\sigma}\left(t - \sum_{i=1}^{d_v}\mu_i\right)\right] \tag{2.3}$$

Since Z has standard normal distribution, we can find the maximum value of t, such that the right-hand side of Equation 2.3 is no less than η.

We select 10 as the passing threshold (e.g. if a node has η-*degree* lower-bound greater than or equal to 10, we keep it, otherwise it will be removed). The rationale is if a vertex has at least 10 edges incident to it before peeling, we can consider it a hotspot suited for the afterward high activity subgraph mining. If in the full network a node is not connected to at least 10 other nodes, there is no point in performing core decomposition as we only focus on dense sub-communities. Procedure for the second data screening stage is described in Algorithm 1.

Algorithm 1 Selection based on η-*degree* lower-bounds

```
1: procedure SecondStageScreening ()
2:     nodelist ← list of remaining nodes in network
3:     init_η_degree ← {}                                  ▷ empty hash table
4:     for all v ∈ nodelist do
5:         init_η_degree[v] ← compute initial η-deg(v)
6:     for all v ∈ init_η_degree.keys() do
7:         if init_η_degree[v] < threshold then
8:             delete init_η_degree[v]
       return init_η_degree.keys()                         ▷ return hash table keys
```

Note that as we start the peeling process, the node's η-*degree* will also start decreasing, so in this last data filtering stage, we only select nodes based on their initial η-*degree* lower-bounds.

There are many nodes in Biomine that can be directly connected to the PPI network, which are potentially more useful than other nodes. In the nodes retaining step, we force them to not be screened out and let downstream analyses decide whether they are useful.

In the merged dataset, 47040 unique Biomine nodes are directly connected to the 692 proteins in the PPI network (a total of 47732 nodes). To preserve valuable information, we retain them from data screening. All other nodes in the *human* Biomine database that are not directly connected to the PPI network will be subject to the two data screening steps.

After this step, a total of 57302 nodes remained in our dataset and the filtered dataset was passed to PA for probabilistic core decomposition with η set to 0.5.

2.4.2 Step 2: Peeling algorithm to find coreness of nodes

In this section, we briefly describe the graph peeling algorithm (PA) introduced by [32].

Core Decomposition based on peeling algorithm includes three important steps: (1) removing vertex u of the smallest η-degree, (2) assigning the core number of u to be equal to its η-degree, and (3) recomputing the η-degree of u's neighbors. Vertices should be kept sorted by their current η-degree at all times during the process. This process is challenging in probabilistic graphs as it involves many recomputations of η-degrees. It should be noted that computing η-degree of a vertex u using dynamic programming takes $O(d_u^2)$. As a result, in [32], an efficient version of the peeling algorithm is proposed which uses efficient array structures and lazy updates of η-degree of vertices.

As mentioned before, η-$deg(v)$ is defined as the highest t for which $\Pr[deg_{\mathcal{G}}(v) \geq t] \geq \eta$. In [32], Lyapunov Central Limit Theorem (CLT) is applied to approximate $\Pr[deg_{\mathcal{G}}(v) \geq t]$ and find the largest t such that $\Pr[deg_{\mathcal{G}}(v) \geq t] \geq \eta$. They showed that Lyapunov CLT can produce a very accurate lower-bound on vertex's true η-*degree*. Since the lower-bound is easy to compute, it helps reduce PA's running time significantly.

To summarize PA, it first computes the lower-bound on the η-*degree* for each

vertex using Lyapunov CLT, then it stores vertices in an array in ascending order of their lower-bound values. Then, the algorithm starts processing vertices based on their (lower-bound on) η-*degree*. When a vertex v is being processed, PA algorithm determines whether v's exact η-*degree* is available or v is on its lower-bound. If the former criteria holds, PA sets v's coreness to be equal to its η-*degree* at the time of process, removes v, and decreases the η-*degree* (exact or lower-bound) of v's neighbours by one. Otherwise, if v is on its lower-bound, v's exact η-*degree* is computed and v is swapped to the proper place in the array. At the end, (k, η)-core of $\mathcal{G}$ is obtained by collecting all the vertices with coreness at least k. Note that for efficiency reasons, (1) lower-bounds are used in the main parts of PA where η-*degree* values are required, and only when a vertex becomes a candidate for removal the exact η-*degree* is calculated, (2) after removing a vertex v, the step of updating the η-*degree* of v's neighbours is delayed as much as possible (*lazy updates* strategy).

2.4.3 Step 3: Functional enrichment analysis

Pathway analysis could prove crucial in understanding how the virus infects the human body [11]. To evaluate functional pathways of proteins involved in SARS-CoV-2 host interactions from the core decomposition result of PA, gene enrichment analysis was performed using clusterProfiler [80] and Metascape [87]. P-values were calculated by hypergeometric test, adjusted using Benjamini–Hochberg procedure, and adjusted p-value < 0.01 were used as the threshold of significance to control the false-discovery rate. We also performed DAVID functional annotation clustering [36, 64] on selected subgraphs. Since Metascape and DAVID both restrict input gene list size up to 3000, if our list exceeds that number, we will select the top 3000 nodes based on their connections to the original PPI network nodes. For uniprot indexing nodes, we select UNIPROT_ACCESSION as the gene list identifier. Everything else is kept as default.

Chapter 3

Results

We merged the original SARS-CoV-2-host protein-protein interaction (PPI) network with the *human* Biomine database and we removed duplicated and looped edges in the merged dataset. We then passed the merged dataset through our proposed analysis pipeline. Approximately 6.6% of nodes remained after data screening and the algorithm revealed the presence of 88 cores. More specifically, the 57302 nodes that remained in the filtered dataset yielded 79 different coreness values ranging from 0 to 88. Many node types exist in the filtered dataset. Some are indexes, for example, UniProt, STRING, PubMed, GO (including indexes for biological process, cellular component, molecular function), etc. Approximately 29.33% of nodes were assigned coreness 1 and 2, which accounts for 13297 nodes and 3507 nodes, respectively. The nodes in the original SARS-CoV-2-human protein-protein interaction (PPI) network were distributed across 73 different coreness with minimum coreness of 1 and maximum coreness of 88. The nodes that were directly connected to the original PPI network nodes (we refer to them as *level 1 connections*) had similar node count distribution with roughly 35.66% nodes assigned with coreness 1 and 2, followed by coreness 77 that contains 2159 ($\approx 4.59\%$) nodes.

Table 3.1 shows the top-10 coreness values in terms of node count for the three scenarios: original SARS-CoV-2-host PPI network nodes, level 1 connections, and complete nodes set. As can be seen from Table 3.1, core 69 and core 77 are the two denser cores that contain a significant fraction of nodes. Compared with other cores in the table (e.g. the core number is less than 30 in the top-10 ranking), core 69 and 77 showed denser sub-communities, indicating that they might contain more valuable information than the others. As indicated before in Equation 2.1, the denser cores are at the same time among the smaller cores, so we further merged all cores

Table 3.1: Top-10 coreness values revealed by peeling algorithm (sorted by node count)

PPI network nodes (coreness(count))	Level 1 connections (coreness(count))	All nodes (coreness(count))
21(34)	1(13288)	1(13297)
23(34)	2(3488)	2(3507)
15(31)	77(2159)	77(2179)
18(26)	3(1545)	9(2085)
69(24)	4(989)	10(2038)
22(23)	23(923)	11(1759)
17(22)	69(872)	12(1584)
14(22)	10(836)	3(1568)
25(21)	15(813)	14(1507)
12(21)	9(794)	8(1499)
77(20)	-	-
19(17)	-	-
13(17)	-	-
11(16)	-	-
26(16)	-	-
24(16)	-	-

In the table, for duplicated node counts, we display all matching coreness.

that are denser than 68 (i.e. coreness 69, 70, 71, 72, 73, 74, 75, 76, 77, 88) into a giant subgraph to avoid losing potential connections as well as the corresponding information between cores.

The resulting subgraph contains 3615 nodes and 2018586 edges. By definition, the merged subgraph is the same as core 69 (it contains nodes with coreness 69 and above). Among these 3615 nodes, a total of 54 nodes are SARS-CoV-2 interacting human proteins identified by Gordon et al. [34] ($\approx 7.8\%$) and the other 3561 nodes are all level 1 connections. In addition, a majority of nodes (3475, $\approx 96.13\%$) in the subgraph are proteins labeled with corresponding UniProt ID. A complete dissection of node types for the subgraph can be found in Table A.1

3.1 SARS-CoV-2 Associating Genes Discovery

We hypothesized that other protein nodes connected with the reported 54 SARS-CoV-2 interacting human proteins in different cores within the subgraph may contain potential missing connections from the single experiment, and provide novel molecu-

lar components for better understanding the pathogenicity of SARS-CoV-2 infection which will eventually be beneficial for identifying new biological/pharmaceutical targets.

We first explored the distribution of the 54 SARS-CoV-2 interacting human proteins in the merged subgraph, and observed that 24 of them has coreness 69, 2 of them has coreness 71, 1 of them has coreness 73, 20 of them has coreness 77, and 7 belongs to core 88 with coreness 88 assigned (Figure 3.2). Since our goal is to extend the SARS-CoV-2-human protein-protein interaction (PPI) network found by Gordon et al. and generate more research hypotheses, in the merged subgraph (the extended network), all the nodes not belonging to the original PPI network can be considered as generated hypotheses. We then try to identify the hypotheses that had already been studied by other researchers. From the subgraph, we obtained 3421 uniprot indexing nodes that are directly connected to the 54 SARS-CoV-2 interacting proteins. We further ranked these 3421 nodes by their connections to the reported proteins (i.e. *nConnect*) and performed literature mining on their associations/correlations with COVID-19. After an automatic web crawling followed by manual inspection, we identified 314 nodes ($\approx 9.18\%$) that show associations or correlations with COVID-19 supported by at least one study. Among the 314 literature verified nodes, 68 of them has coreness 69, 198 of them has coreness 77 (in total, $\approx 84.7\%$ of literature verified nodes were either assigned coreness 69 or coreness 77, a complete plot on coreness distribution is shown in Figure 3.1). As mentioned before, core 69 and 77 showed denser sub-communities compares with others in different contexts in Table 3.1. And for the 54 SARS-CoV-2 interacting human proteins in core 69, 44 of them ($\approx 81.5\%$) either have coreness 69 or coreness 77. We believe these findings support our assumption that more valuable information can be found in dense sub-communities and we move on to explore the criteria for high quality hypothesis (i.e. hypothesis that most likely to be true).

Table A.2 lists the top 55 genes and encoded proteins that are connected to more than 30 ($30/54 = 55.6\%$) virus-interacting proteins in the subgraph, ranked by *nConnect*. The number 55 is not arbitrary, it is simply because there are many ties in the ranking for nodes with $nConnect \geq 30$. Over 30% of nodes in Table A.2 received literature supports. Therefore, we consider nodes in Table A.2 to be high quality hypotheses compared to the rest of the nodes in the subgraph and we recommend researchers start validating them first. The subgraph node list and the complete list of nodes that received literature supports can be found in Supplementary Tables

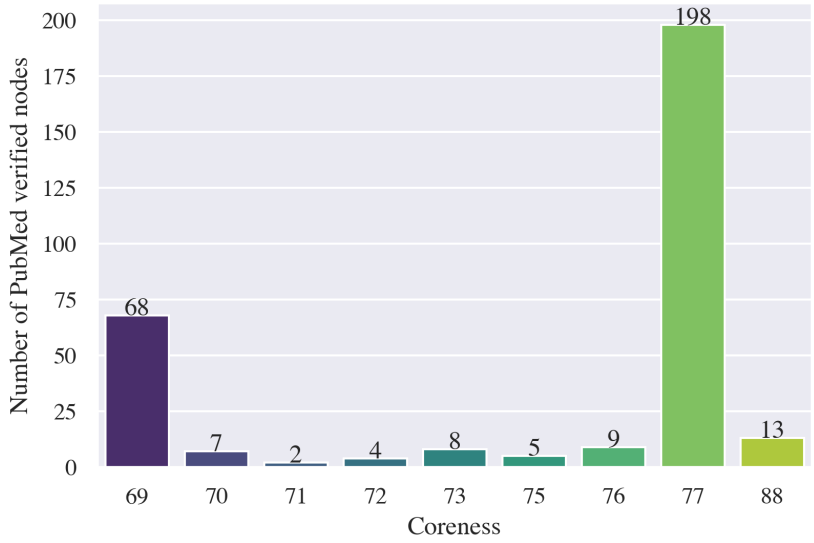


Figure 3.1: Distribution of coreness in the list of PubMed verified nodes

1 and 2 in Appendix. The list of all potential genes and proteins that related to COVID-19 in the subgraph (including the aforementioned high quality hypotheses) can be found in Supplementary Table 5 in Appendix. A detailed discussion on the literature-supported nodes in Table A.2 can be found in the Discussion section.

3.2 Gene Ontology Over-representation Analysis

We first performed an enrichment test of gene ontology (GO) biological process for genes in the subgraph. The top 30 GO terms with the smallest p-value were presented in Figure 3.3. The most significant GO term is *protein polyubiquitination* (adjusted p-value $\approx 5.1176 * 10^{-75}$), which accounts for 204 of the total 3342 mapped input genes ($\approx 6.1\%$), followed by *ribonucleoprotein complex biogenesis* (adjusted p-value $\approx 6.8251 * 10^{-58}$). The top enriched term might suggest that SARS-CoV-2 hijacks cell's ubiquitination pathways for replication and pathogenesis, which is one of the findings of [34]. Interestingly, the virus-associated biological processes *viral gene expression* and *viral transcription* were also found to be enriched and account for about 3.7% and 3.4% of input eligible gene set, respectively. In addition, we noted that 10 out of the top 30 GO terms were RNA-related. For example, *mRNA catabolic process* (6% of total input eligible genes), *RNA catabolic process* (6.3%), and *RNA*

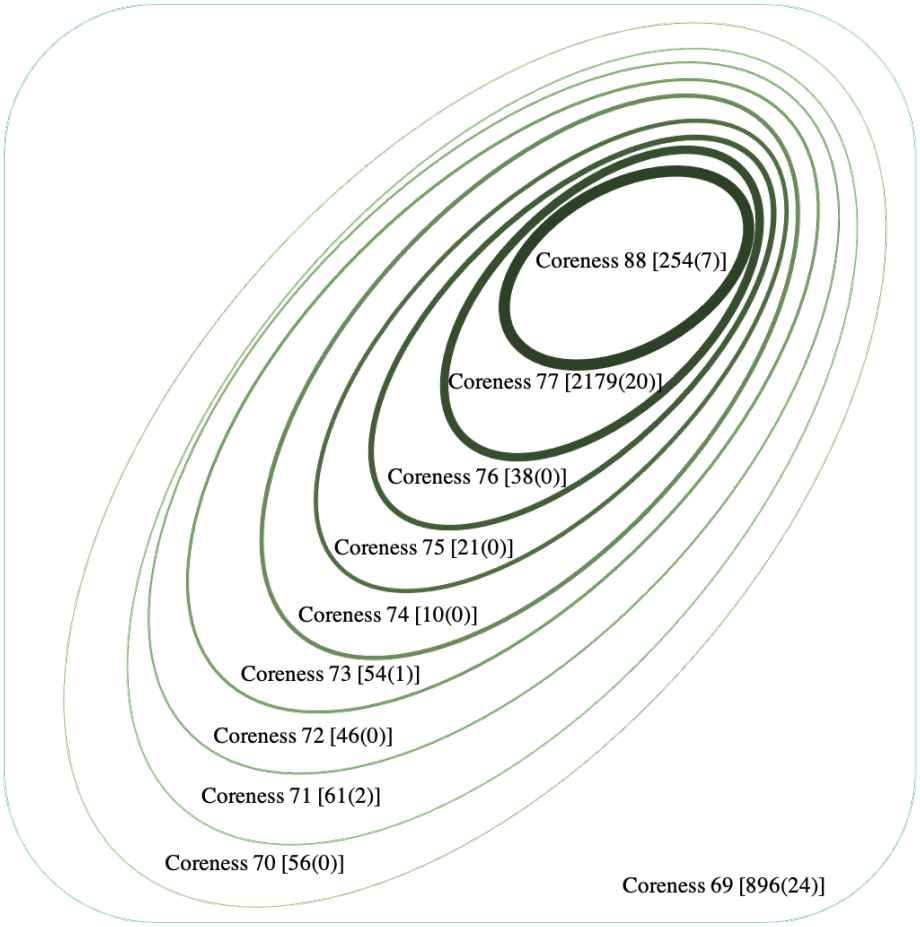


Figure 3.2: Illustration of the merged subgraph, the number in square brackets represents the number of nodes with the specified coreness while the number in parentheses represents the number of PPI network nodes with that coreness

splicing (6.2%) were the top items. This is in line with Gordon's study where they found SARS-CoV-2 proteins NSP8 and N involved in RNA processing and regulation [34].

We further used Metascape to perform pathway and process enrichment analysis on uniprot indexing nodes with different coreness. Two subsets showed significant enrichment in the molecular functions of immune response to bacteria or viruses. Specifically, 5.77% protein-encoding genes with coreness 70 enriched the term *The human immune response to tuberculosis* and other 5.77% genes enriched the term *regulation of viral process*, see Figure 3.4 for other relevant enriched items (*negative regulation of protein kinase activity*, *Signaling by Receptor Tyrosine Kinases*, etc.). Regarding uniprot indexing nodes with coreness 73, Blanco-Melo et al. [10] found

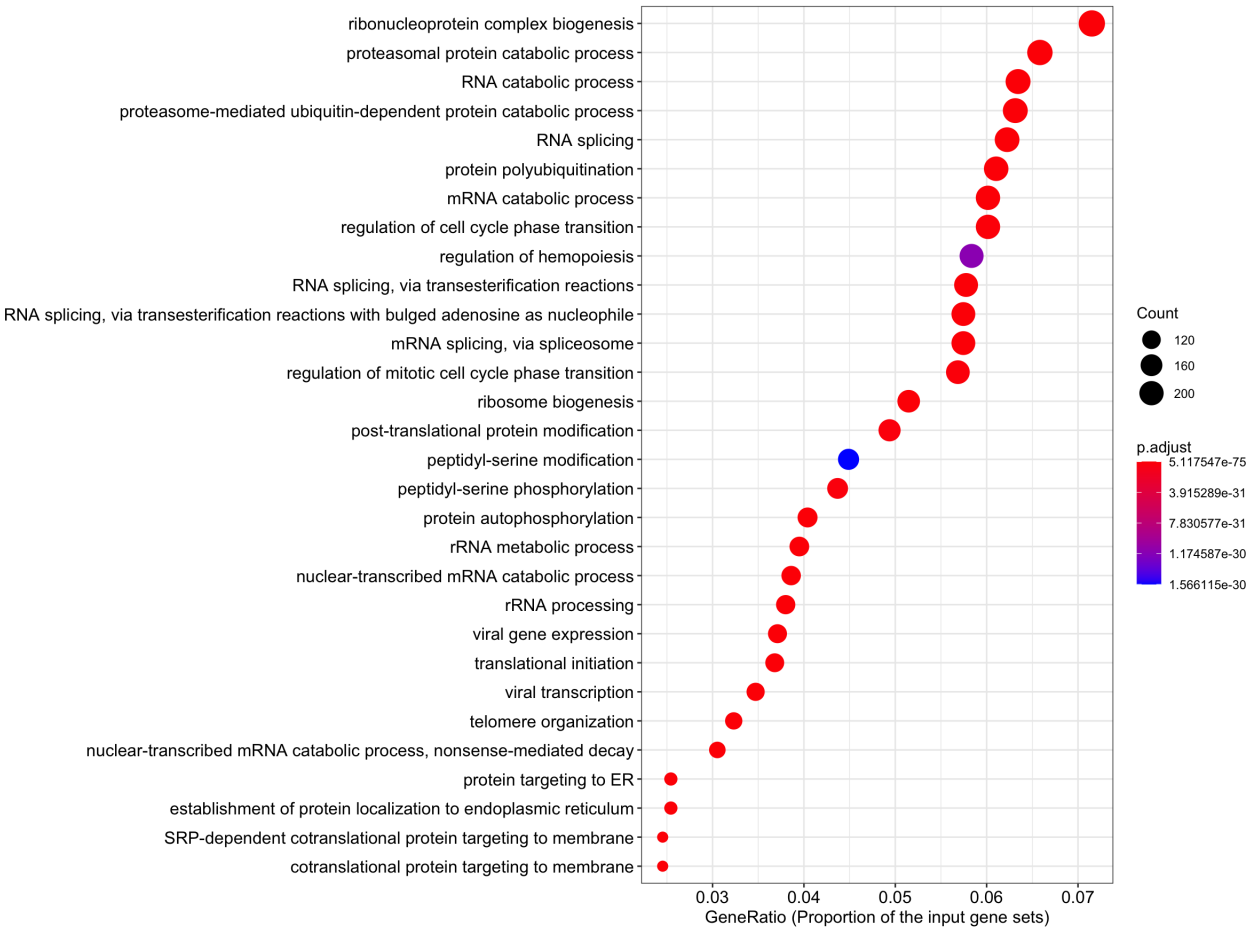


Figure 3.3: Top GO terms ranked by GeneRatio

the excessive expression of cytokine as one of the strong features of SARS-CoV-2 infection, here we are able to find several terms that are related to cytokine storm induced by SARS-CoV-2 (see Figure 3.5). For instance, *IL-4 Signaling Pathway* and *Cytokine Signaling in Immune system* are related to cytokine storm upon virus infection. Particularly, IL-4 is one kind of cytokine that acts as a regulator of the JAK-STAT pathway and contributes to human body immune responses [56].

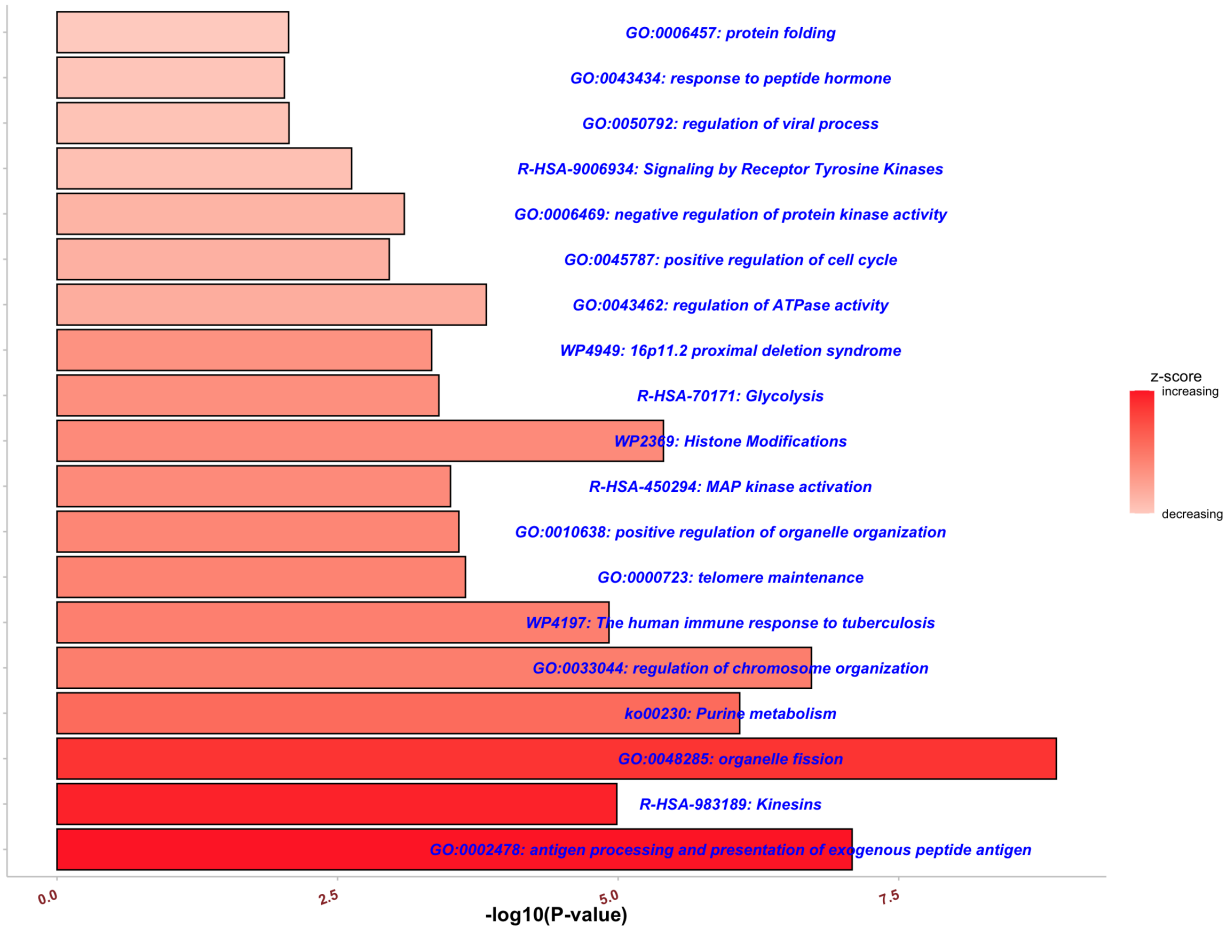


Figure 3.4: Top biological terms (pathway, process, etc.) enriched for coreness 70 ranked by p-values

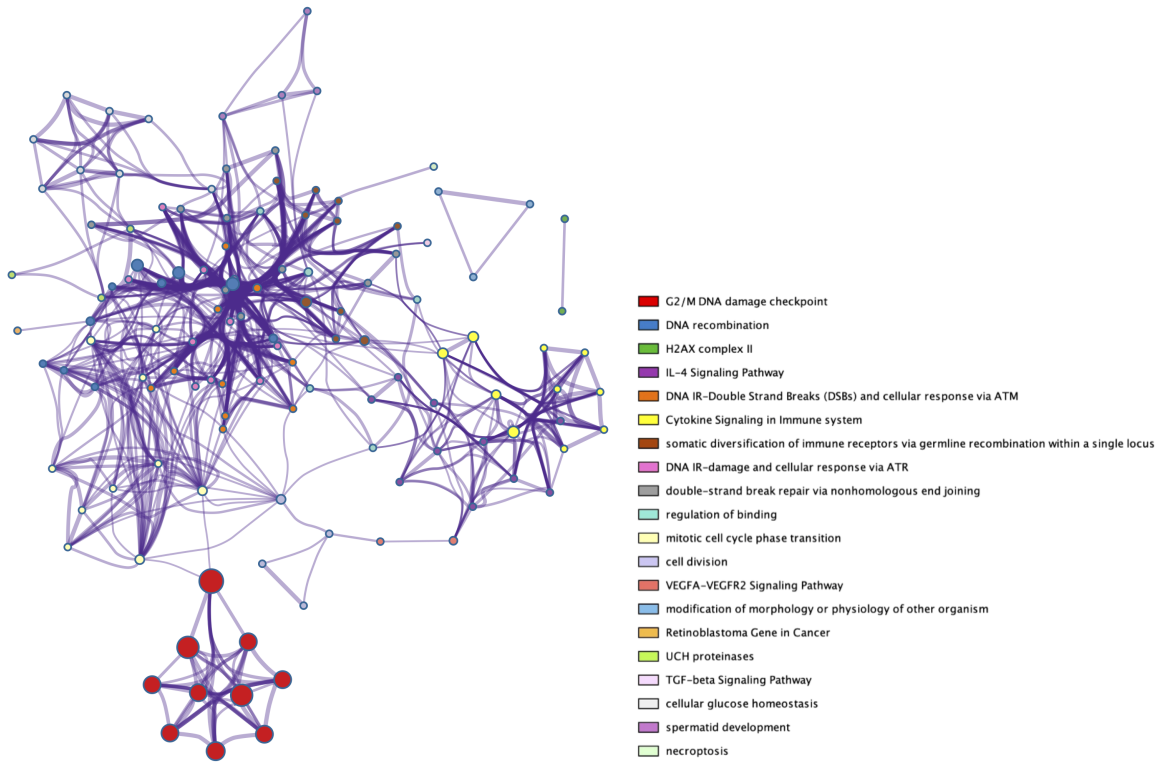


Figure 3.5: Network of enriched terms for coreness 73 colored by cluster ID, nodes shared the same cluster ID typically lie close together

3.3 SARS-CoV-2 Interacts With Tyrosine-related Proteins

Bouhaddou et al. [15] found changes in activities for 97 out of 518 human kinases during SARS-CoV-2 virus infection. Surprisingly, among the 97 kinases list they discovered, 73 were found in our merged subgraph ($\approx$ 75.3%). This motivated us to check for all the 302 kinases-related uniprot indexing nodes in our subgraph (the complete list of kinases-related nodes can be found in Supplementary Table 3 in Appendix). Interestingly, 48 of them are tyrosine-related nodes ($\approx$ 15.9%). As mentioned previously, we found 314 out of the 3421 uniprot indexing nodes ($\approx$ 9.18%) in the merged subgraph that had at least one study showing they have some relations with COVID-19. When we restrict to tyrosine-related proteins, this proportion increased fourfold to almost 40% (Table 3.2). That is, 24 of the 61 tyrosine-related proteins (Supplementary Table 4 in Appendix) in the merged subgraph have been reported to be associated with SARS-CoV-2 virus, including SRC, TYRO3, FLT3, TYK2, LYN, BTK, LCK, SYK, BLK, ERBB2, etc. The high proportion of validated tyrosine-related proteins further motivated us to perform a DAVID functional annotation clustering on uniprot indexing nodes in core 69 (the merged subgraph). In total, 336 clusters were identified and the term *Tyrosine-protein kinase* (fold-change = 5, p-value = $3.5 * 10^{-42}$) is among the top 15 clusters ranked by enrichment score (the DAVID analysis report can be found in Supplementary Report 1 in Appendix). Of particular importance, SRC, JAK1, JAK2, ABL1 are among the top of the list with large connections to the original SARS-CoV-2 interacting tyrosine-protein kinase network. Figure 3.6 presented a complete PPI network between tyrosine-protein kinase SRC, JAK1/2, ABL1/2 and their connections to the 54 SARS-CoV-2 interacting human proteins in the subgraph. We also include the SARS-CoV-2 viral proteins (red diamond) from the original PPI network. This network covered half (n=13) of the 26 SARS-CoV-2 viral proteins, including envelope (E), NSP7, NSP10, ORF10, etc. It showed these four proteins were not directly interacting with the virus but at level one connections where SRC, JAK1/2, and ABL1/2 are hub nodes (connection degree being 41, 37, 36, 35, and 29, respectively).

Additionally, we ranked all the nodes in the merged subgraph by their degrees and investigated the top 5% of the sorted nodes (181 out of 3615 nodes, Supplementary Table 1). There are 79 "kinase" protein nodes and 28 of them are "tyrosine kinase". Among the 28 tyrosine-related nodes, the top ranking ones were SRC (ranked 5 out

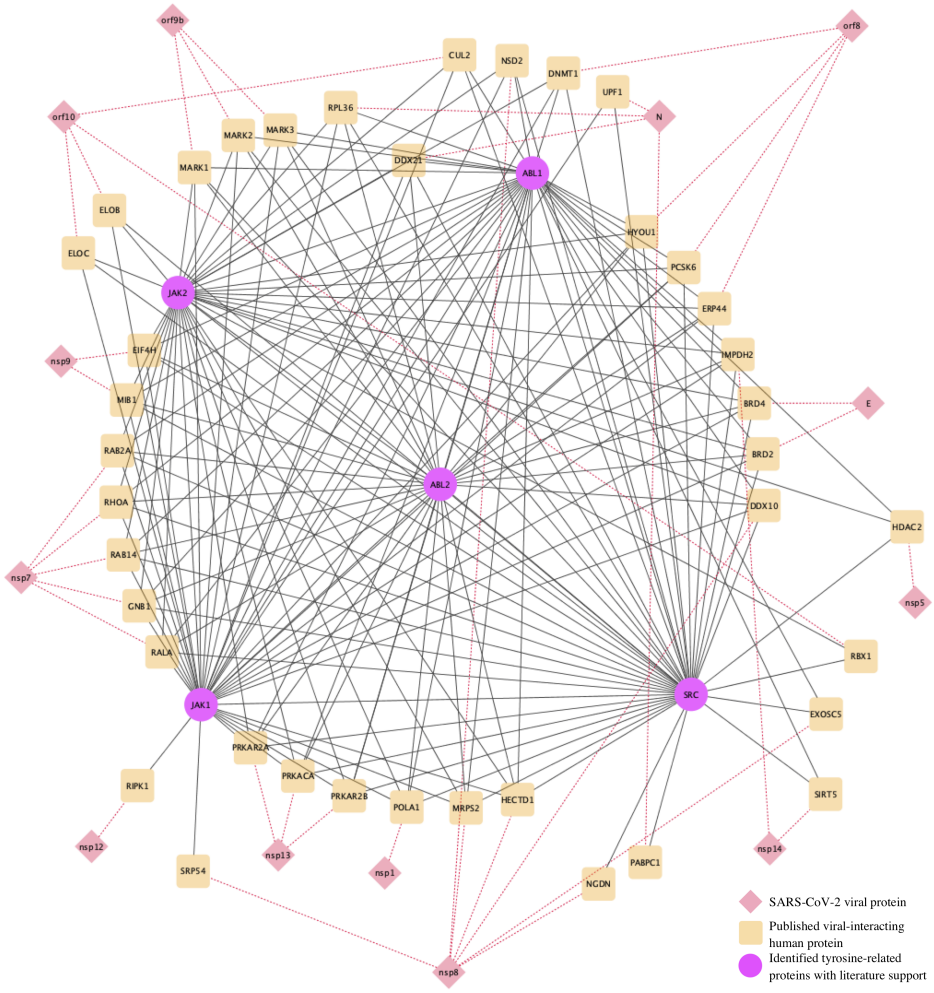


Figure 3.6: SARS-CoV-2 protein interaction with Tyrosine-protein kinase JAK1/2, ABL1/2, SRC. Dashed edges indicate the proteins do not belong to the identified subgraph.

of 181), ABL1 (ranked 20 out of 181), and JAK2 (ranked 21 out of 181) with degree connection 2682, 2479, and 2476, respectively. Take together, evidence obtained so far highly implied that SRC, ABL1, and JAK2 are three important hub genes.

3.4 COVID-19 Related Tissues and Diseases Discovery

Through the 332 human interacting proteins across different tissues, Gordon et al. [34] identified lung as the tissue with the highest level of differential expression of the

Table 3.2: All tyrosine-related proteins in the subgraph that received literature support

UniProt ID	Gene Name	Protein Name
P12931	SRC	Proto-oncogene tyrosine-protein kinase Src
P23458	JAK1	Tyrosine-protein kinase JAK1
O60674	JAK2	Tyrosine-protein kinase JAK2
P00519	ABL1	Tyrosine-protein kinase ABL1
Q06418	TYRO3	Tyrosine-protein kinase receptor TYRO3
P36888	FLT3	Receptor-type tyrosine-protein kinase FLT3
P29597	TYK2	Non-receptor tyrosine-protein kinase TYK2
P09769	FGR	Tyrosine-protein kinase Fgr
P07948	LYN	Tyrosine-protein kinase Lyn
Q06187	BTK	Tyrosine-protein kinase BTK
P06239	LCK	Tyrosine-protein kinase Lck
P42684	ABL2	Tyrosine-protein kinase ABL2
P42680	TEC	Tyrosine-protein kinase Tec
P43405	SYK	Tyrosine-protein kinase SYK
P51451	BLK	Tyrosine-protein kinase Blk
Q08881	ITK	Tyrosine-protein kinase ITK/TSK
P08631	HCK	Tyrosine-protein kinase HCK
P41240	CSK	Tyrosine-protein kinase CSK
P30530	AXL	Tyrosine-protein kinase receptor UFO
Q12866	MERTK	Tyrosine-protein kinase Mer
P04626	ERBB2	Receptor tyrosine-protein kinase erbB-2
P08575	PTPRC	Receptor-type tyrosine-protein phosphatase C
Q9Y463	DYRK1B	Dual specificity tyrosine-phosphorylation-regulated kinase 1B
P17706	PTPN2	Tyrosine-protein phosphatase non-receptor type 2

SARS-CoV-2 interacting human proteins. They also found 15 other tissues with a high abundance of SARS-CoV-2 human interacting proteins. In our subgraph, we were able to locate 5 out of the top 16 tissues identified by [34]: lung (coreness=69), testis (coreness=69), placenta (coreness=69), liver (coreness=69), and brain (coreness=77). Curiously, we also discovered two diseases that seem to be associated with COVID-19 within the subgraph (Supplementary Table 5): cervix carcinoma (coreness=70) and erythroleukemia (coreness=69). Figure 3.7 presents a network consisting of the 5 identified tissue nodes, 2 disease nodes, and 2978 uniprot indexing nodes that directly connect to tissue nodes and disease nodes in the subgraph (an interactive version of the network can be found in Supplementary Material 1 in Appendix). 51 out of the 54 reported SARS-CoV-2 interacting human proteins can be found in this

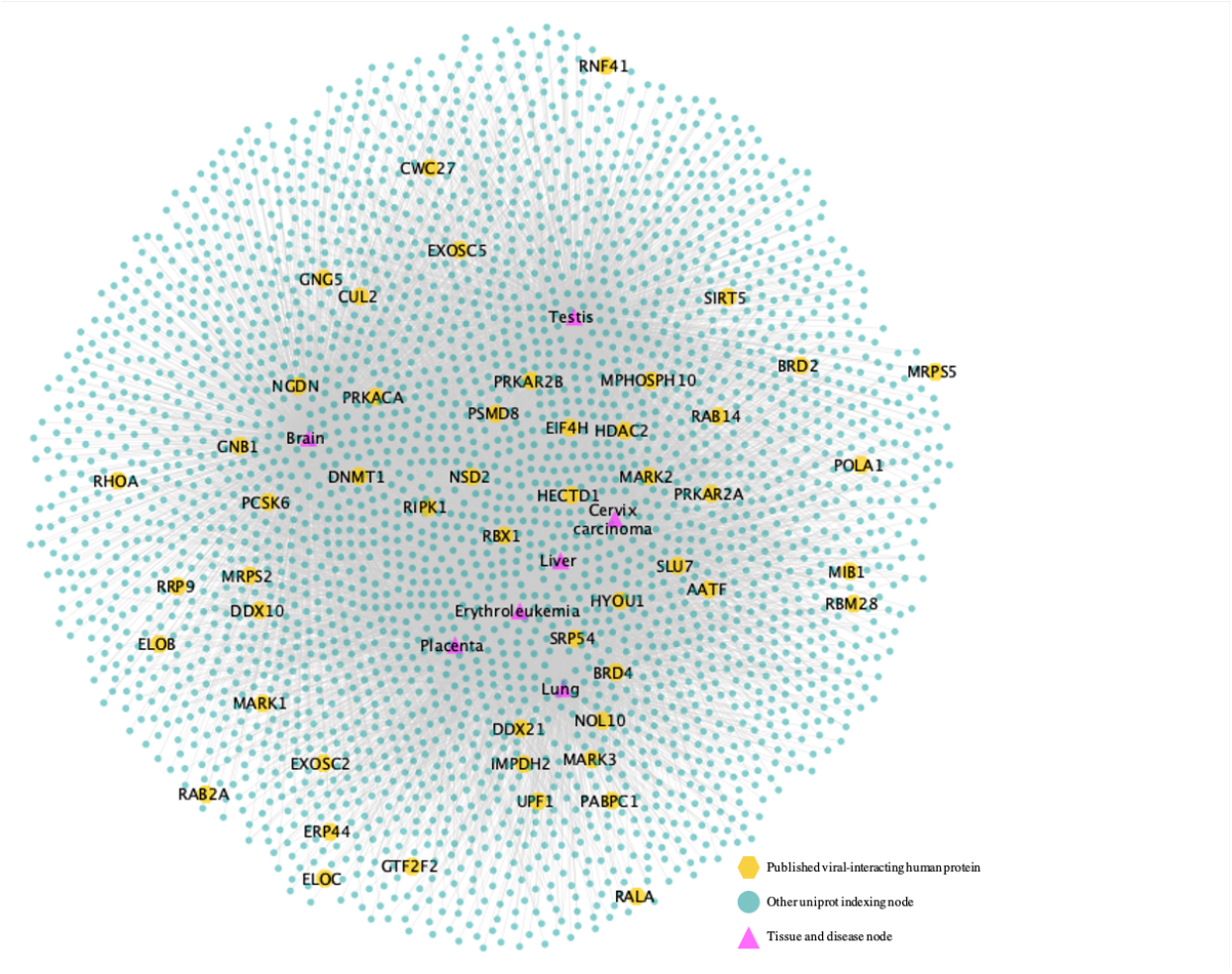


Figure 3.7: Interaction map between tissue, disease, and uniprot indexing nodes in the subgraph (all of the edges have is_expressed_in as edge relationship type)

network. Figure 3.8 shows the interaction between the 51 SARS-CoV-2 interacting human proteins and the tissue and disease nodes identified in the subgraph under a higher resolution. Similar to what Gordon et al. [34] found, lung, testis, placenta, liver, and brain are heavily involved during SARS-CoV-2 infection. For example, the HDAC2 protein, which is observed to be expressed in testis, lung, overexpressed in cervical cancer, etc., has been identified by Gordon et al. to be associated with the SARS-CoV-2 NSP5 protein. We noted that the same SARS-CoV-2 interacting human proteins are also expressed in cervix carcinoma and erythroleukemia in the sub-graph. We hypothesized drugs used to treat these two diseases might be useful to treat COVID-19.

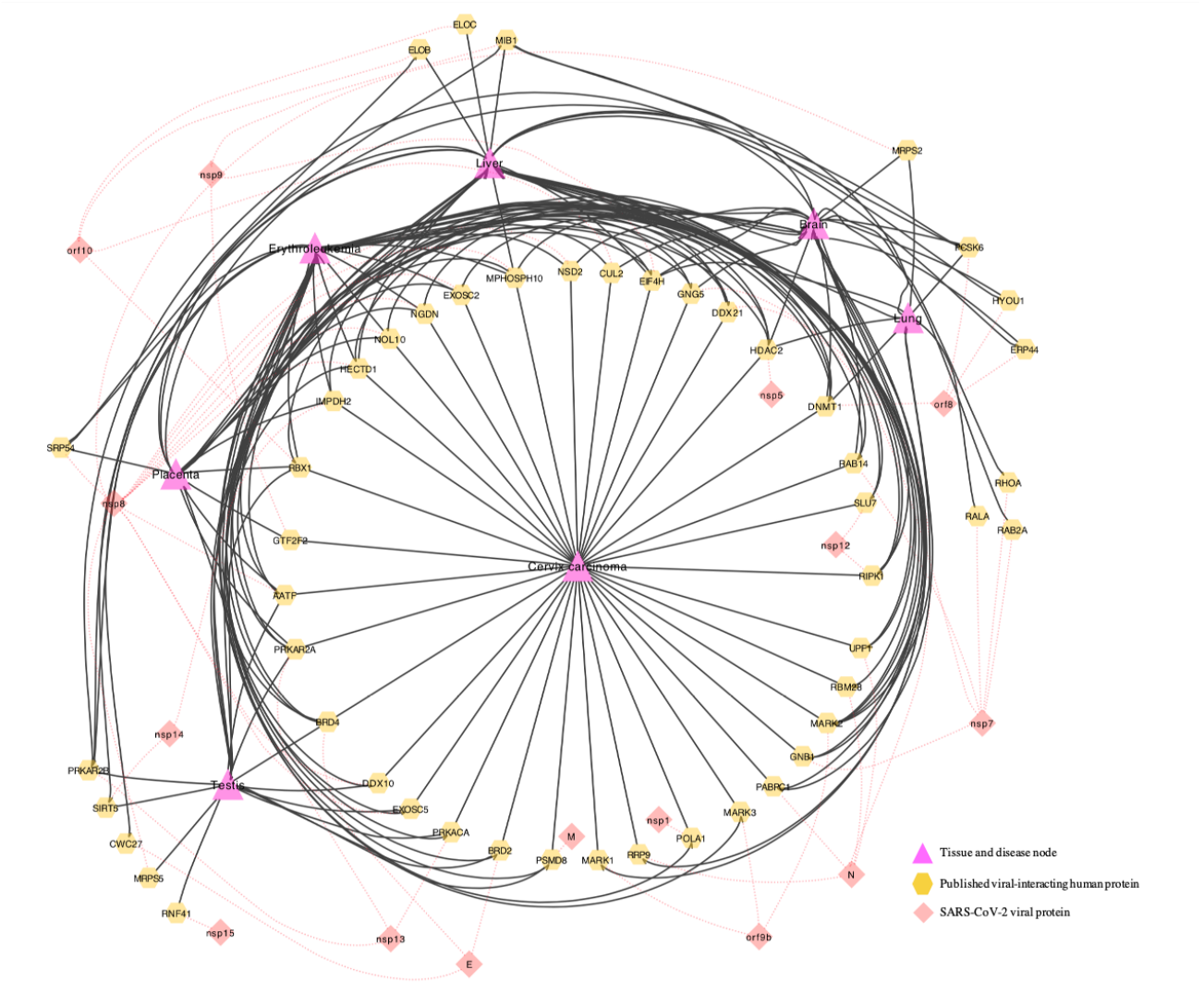


Figure 3.8: Association between SARS-CoV-2 host interacting proteins and tissue/disease nodes in core 69, all edges have relationship type is_expressed_in, we add SARS-CoV-2 viral proteins (red diamond with dashed links, indicating they do not exist in the subgraph) for clarity

Idarubicin, daunorubicin, and cytarabine are the three drugs used to treat erythroleukemia. Chandra et al. [22] suggest idarubicin as a potential drug that can be repurposed for controlling SARS-CoV-2 infection due to its good binding affinity to SARS-CoV-2 NSP15 encode endoribonuclease enzyme.

Bevacizumab is used to treat cervix carcinoma. According to Rosa et al. [59], there have been clinical trials about repositioning bevacizumab for COVID-19. In

addition, Amawi et al. and Zhang et al. [3, 81] identified bevacizumab as one of the promising therapeutic treatments against COVID-19.

Chapter 4

Discussions

Using the rich Biomine database, we extended the SARS-CoV-2-human protein-protein interaction (PPI) network identified by Gordon et al. [34]. Through the proposed three-stage analysis pipeline, we were able to filter the large extended network, uncover dense sub-communities, and therefore generate research hypotheses related to COVID-19 by identifying dense cores in the Biomine database that have as many same nodes as the PPI network.

We assumed more connections (more activities) in a subgraph means that it contains more interesting information. In the first step of our proposed pipeline, we have two data-screening sub-steps added; this design is based on our assumption and will enable the peeling algorithm (PA) to focus more on cores with high activities. Furthermore, by focusing on dense cores, the PA can run faster due to a decrease in the input dataset size.

It is worth noting that when the network contains more than one connected component after the two-stage data screening, our proposed pipeline will produce the result even faster since the network complexity will be further reduced and we can run PA in parallel for each connected component. But in Biomine, the network is connected after screening and we were unable to run PA in parallel.

Based on the aforementioned assumption, the Biomine database, and the proposed analysis pipeline, we quickly identified sub-communities in the extended PPI network that have high activities and we discovered novel diseases, genes, and proteins that could potentially relate to COVID-19 as research hypotheses. The generated hypotheses (Supplementary Table 5 in Appendix) provide candidates for follow-up work to validate.

Kumar et al. [44] also employed a similar graph decomposition concept on a host-

viral network. The difference between their analysis approach and ours is that they started with a deterministic graph and added edge weight later while we started with probabilistic graph decomposition which is more challenging. Another notable difference is they utilized a graph decomposition method called weighted k-shell. In contrast to our k-core decomposition, k-shell asks for all the nodes in the subgraph to have coreness precisely equal to k while k-core asks for all nodes' coreness to be at least k. k-core has the advantage of connecting different subsets of nodes in the network and helps discover missing links between them. Additionally, when we perform pathway and process enrichment analysis on the subgraph, we also use the subset of nodes with the same coreness, which by definition is the k-shell of the subgraph.

For determining the merged subgraph, we need to pick an optimal threshold of coreness that will balance the denseness and the complexity of the graph. We observed that core 69 and 77 are both dense and contains more information. Compared to core 77, we preferred core 69 as we want to include more nodes into the subgraph. In another hand, it is not practical to involve too many nodes which results in generating a vast amount of hypotheses to validate. Therefore after comparing with coreness 77 and others with coreness around 20 in Table 3.1, we chose coreness 69 as the final optimal threshold.

We did a literature review for most of the genes detected in our subgraphs and listed genes with evidence in Table A.2. Take entries with $nConnect \geq 34$ as an example, 5 genes received literature support. The rest 8 genes in that subgraph might be strong candidates with high priority and worth for future validation. For GAPDH (Glyceraldehyde-3-phosphate dehydrogenase), Zheng et al. [84] obtained potential COVID-19 effector targets of Chinese medicine Xuebijing (XBJ) which was used in treating mild cases of COVID-19 patients, and GAPDH is found to be one of the key targets. Taniguchi-Ponciano et al. [67] identified HIF1α as a potential marker for COVID-19 severity and GAPDH is found to be among the expressed HIF1α responsive genes. Ebrahimi et al. [29] found inhibiting GAPDH in patients with degenerated innate immunity can potentially help in treating COVID-19. For SRC (Proto-oncogene tyrosine-protein kinase Src), in Lin et al. [45], ibrutinib is found to block SRC family kinases, which might reduce viral entry as well as the inflammatory cytokine response in the lungs. Morenikeji et al. [49] identified SRC to be one of the genes associated with Bovine coronavirus and by implication, other coronaviruses. Tiwari et al. [68] found SRC to play a vital role in SARS-CoV-2 infection

related pathways. Xie et al. [78] found SRC participates in cytokines storm in patients with obesity which could lead to negative outcomes when infected with SARS-CoV-2. For UBC (Polyubiquitin-C), it is found to be one of the novel host factors predicted to impact the replication cycles of SARS-CoV [28]. For HSP90AA1 (Heat shock protein HSP 90-alpha), Wicik et al. [73] found it to be among the strongest interaction for ACE2-related co-expression networks. Lastly, for AKT1 (RAC-alpha serine/threonine-protein kinase), it is identified as a potential drug target for treating COVID-19 [77]. In addition, Zhao et al. [83] examined the underlying molecular mechanism of widely used COVID-19 treating Chinese medicine Qingfei Paidu decoction and found AKT1 to be related to its effects. Similarly, Zhuang et al. [90] found AKT1 to be 1 of the 10 hub target genes by Shufeng Jiedu capsule (a popular applied COVID-19 Chinese medicine treatment). Appelberg et al. [4] found AKT1 to be upregulated in SARS-CoV-2 infected Huh7 cells. Besides finding SRC, Tiwari et al. [68] also discovered AKT1 to play a vital role in SARS-CoV-2 infection related pathways.

For tyrosine-related proteins in the merged subgraph and their associations with COVID-19, many tyrosine kinase inhibitors have been identified to inhibit the SARS-CoV-2 virus. Cagno et al. [18] found three ABL tyrosine-protein kinase inhibitors imatinib, dasatinib, and nilotinib to exert inhibitory activity against SARS-CoV-2, Alijotas-Reig et al. [2] found two JAK tyrosine-protein kinase ruxolitinib and baricitinib to be useful in treating the COVID-19 induced systemic hyperinflammatory response (cytokine storm). Wu et al. and Seif et al. [75, 63] found another JAK inhibitor fedratinib to mitigate the serious conditions in COVID-19 patients. For the hub nodes identified in Figure 3.6, we have discussed literature support for SRC before, and will now discuss supports for JAK1/2 and ABL1/2. Many works of literature have targeted JAK1/2 in the hope to treat or prevent COVID-19. Shi et al. [65] found decreasing of lymphocyte in patients with COVID-19 correlated with low expression of JAK1-STAT5 signaling pathway. Zhang et al. [82] found that by suppressing JAK1/2 using baricitinib, several cytokines signals inciting inflammation will be inhibited. Others have also suggested using drugs that inhibit JAK1/2 to help patients with COVID-19 [20, 79, 19, 33, 7, 69, 66, 60]. As for ABL1/2, Abruzzese et al. [1] have suggested a possibility that patients treated with BCR-ABL tyrosine kinase inhibitors may be protected from the virus infection.

Chapter 5

Conclusions

With the SARS-CoV-2 outbreak being declared as a pandemic by the World Health Organization (WHO) [27], researchers around the globe are shifting their focus to COVID-19. In this work, we are especially interested in the protein-protein interaction (PPI) network between SARS-CoV-2 proteins and human proteins identified by Gordon et al. [34], and our focus lies in extending the network to generate biological hypotheses for further validation. To achieve that, we connect the single experiment derived PPI network with the large Biomine database and aim to locate sub-communities with high activities in the extended network. We propose a data analysis pipeline based on a graph peeling algorithm (PA) that enabled us to compute core decomposition efficiently. We select dense cores in the Biomine database that overlap most with the PPI network. The dense subgraph is the resulting extended network and every non-PPI node in the subgraph is a generated hypothesis. We then evaluate the selected subgraph in three contexts: we performed literature validation for uncovered virus targeting genes and proteins and found genes that have already been validated by others on their relationships to COVID-19; we carried out gene ontology over-representation test on the subgraph and found underlying enriched terms related to viral replication, viral pathogenesis, cytokine storm, etc.; we also searched for literature support on the identified tissues and diseases related to COVID-19 and found the possibility of drug repurposing for COVID-19 treatment. To further assign priorities to the generated hypotheses, we sorted all uniprot indexing nodes in the subgraph by their connections to the PPI nodes. The top ranking nodes (Table A.2) in the list have a high proportion of literature validated nodes (for instance, GAPDH, UBC, HSP90AA1, DICER1, GSK3B, HSPA8, and tyrosine-protein kinase SRC, JAK1, JAK2, ABL1, etc.), we deem the rest non-validated nodes in the table

as high quality hypotheses.

Appendix A

Additional Information

A.1 External Repository

Due to limitations in content size, several tables and supporting documents cannot be included even in the Appendix. They are placed in a GitHub repository. Below are specific links to the files:

- Supplementary Report 1.pdf
- Supplementary Material 1.zip
- Supplementary Table 1.xlsx
- Supplementary Table 2.xlsx
- Supplementary Table 3.xlsx
- Supplementary Table 4.xlsx
- Supplementary Table 5.xlsx

Table A.1: Node types in core 69 (sorted by node count)

Node type	Count
UniProt indexing nodes (prefix: 'Protein_UniProt')	3475
STRING indexing nodes (prefix: 'Protein_STRING')	113
PubMed indexing nodes (prefix: 'Article_PubMed')	8
UniProt_Tissue indexing nodes (prefix: 'Tissue_UniProt/tissue')	7
Gene Ontology (GO) Molecular Function indexing nodes (prefix: 'MolecularFunction_GO')	4
GO Cellular Component indexing nodes (prefix: 'CellularComponent_GO')	4
GO Biological Process indexing nodes (prefix: 'BiologicalProcess_GO')	1
InterPro Domain indexing nodes (prefix: 'Domain_InterPro')	1
TAXON Organism indexing nodes (prefix: 'Organism_TAXON')	1
InterPro Homologous Superfamily indexing nodes (prefix: 'HomologousSuperfamily_InterPro')	1

Each type of nodes has specific prefix to its name (listed in parentheses).

Table A.2: Top 55 discovered genes to be potentially related to COVID-19

Gene Name	UniProt ID	*nConnect*	References ('etc.' means found > 10 references)
PRDM10	Q9NQV6	42	-
GAPDH	P04406	40	[84, 67, 29]
RIPK4	P57078	38	-
UBA52	P62987	38	-
SRC	P12931	37	[45, 49, 68, 78]
HSPA4	P34932	37	-
POLR2A	P24928	36	-
EHMT1	Q9H9B1	35	-
EHMT2	Q96KQ7	35	-
UBC	P0CG48	35	[28]
HSP90AA1	P07900	34	[73]
POLR2B	P30876	34	-
AKT1	P31749	34	[77, 83, 4, 90, 68]
RPS27A	P62979	33	-
CAD	P27708	33	-
CDK2	P24941	33	-
PHLPP1	O60346	33	-
DICER1	Q9UPY3	33	[50]
ASH1L	Q9NR48	33	-
POTEF	A5A3E0	33	-
GSK3B	P49841	33	[46, 42, 51, 30]
JAK1	P23458	33	[65, 20, 82, 79, 19, 33, 7, 69, 14], etc.
DDX5	P17844	32	-
HSPA8	P11142	32	[89]
POTEE	Q6S8J3	32	-
EPRS1	P07814	32	-
JAK2	O60674	32	[75, 9, 82, 79, 19, 66, 7, 60, 39], etc.
UBB	P0CG47	32	-
PHLPP2	Q6ZVD8	32	-
UMPS	P11172	32	-
ACLY	P53396	31	-
DHX8	Q14562	31	-
IGF1R	P08069	31	[74]
PPIE	Q9UNP9	31	-
POLR1A	O95602	31	-
YWHAZ	P63104	31	[70]
PIKFYVE	Q9Y2I7	31	[52, 15, 58, 40, 8, 57, 23]
MTOR	P42345	31	[12, 17, 55, 6, 85, 61, 5, 16, 41, 26], etc.
POTEJ	P0CG39	31	-
POTEI	P0CG38	31	-
ABL1	P00519	31	[1]
CDK1	P06493	31	-
PIK3CB	P42338	30	-
CTR9	Q6PD62	30	-
HSPA9	P38646	30	-
MAP3K7	O43318	30	[24]
RET	P07949	30	-
PAICS	P22234	30	-
ACTB	P60709	30	-
PCNA	P12004	30	[48]
HACE1	Q8IYU2	30	-
PIK3CG	P48736	30	-
KIT	P10721	30	-
PIK3CA	P42336	30	-
DDX18	Q9NVP1	30	-

Bibliography

[1] Elisabetta Abruzzese, Luigiana Luciano, Francesco D'Agostino, Malgorzata Monika Trawinska, Fabrizio Pane, and Paolo De Fabritiis. SARS-CoV-2 (COVID-19) and Chronic Myeloid Leukemia (CML): a case report and review of ABL kinase involvement in viral infection. *Mediterranean Journal of Hematology and Infectious Diseases*, 12(1):e2020031, 2020.

[2] Jaume Alijotas-Reig, Enrique Esteve-Valverde, Cristina Belizna, Albert Selva-O'Callaghan, Josep Pardos-Gea, Angela Quintana, Arsene Mekinian, Ariadna Anunciacion-Llunell, and Francesc Miró-Mur. Immunomodulatory therapy for the management of severe covid-19. beyond the anti-viral therapy: A comprehensive review. *Autoimmunity reviews*, 19(7):102569, 2020.

[3] Haneen Amawi, Ghina'a I Abu Deiab, Alaa A A Aljabali, Kamal Dua, and Murtaza M Tambuwala. COVID-19 pandemic: an overview of epidemiology, pathogenesis, diagnostics and potential vaccines and therapeutics. *Therapeutic Delivery*, 11(4):245–268, 2020.

[4] Sofia Appelberg, Soham Gupta, Sara Svensson Akusjärvi, Anoop T Ambikan, Flora Mikaeloff, Elisa Saccon, Ákos Végvári, Rui Benfeitas, Maike Sperk, Marie Ståhlberg, et al. Dysregulation in akt/mtor/hif-1 signaling identified by proteo-transcriptomics of sars-cov-2 infected cells. *Emerging microbes & infections*, 9(1):1748–1760, 2020.

[5] Prabhu S Arunachalam, Florian Wimmers, Chris Ka Pun Mok, Ranawaka APM Perera, Madeleine Scott, Thomas Hagan, Natalia Sigal, Yupeng Feng, Laurel Bristow, Owen Tak-Yin Tsang, et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science*, 369(6508):1210–1220, 2020.

[6] William S Azar, Rachel Njeim, Angie H Fares, Nadim S Azar, Sami T Azar, Mazen El Sayed, and Assaad A Eid. COVID-19 and diabetes mellitus: how one pandemic worsens the other. *Reviews in Endocrine and Metabolic Disorders*, 21(4):451–463, 2020.

[7] Bakiye Goker Bagca and Cigir Biray Avci. The potential of JAK/STAT pathway inhibition by ruxolitinib in the treatment of COVID-19. *Cytokine & growth factor reviews*, 54:51, 2020.

[8] Maksim V Baranov, Frans Bianchi, and Geert van den Bogaart. The PIKfyve Inhibitor Apilimod: A Double-Edged Sword against COVID-19. *Cells*, 10(1):30, 2021.

[9] Andresa Aparecida Berretta, Marcelo Augusto Duarte Silveira, José Manuel Cóndor Capcha, and David De Jong. Propolis and its potential against SARS-CoV-2 infection mechanisms and COVID-19 disease. *Biomedicine & Pharmacotherapy*, 131:110622, 2020.

[10] Daniel Blanco-Melo, Benjamin E Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X Jordan, Kohei Oishi, Maryline Panis, David Sachs, et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, 181(5):1036–1045, 2020.

[11] Denisa Bojkova, Kevin Klann, Benjamin Koch, Marek Widera, David Krause, Sandra Ciesek, Jindrich Cinatl, and Christian Münch. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*, 583(7816):469–472, 2020.

[12] Alireza Bolourian and Zahra Mojtahedi. Obesity and COVID-19: The mTOR pathway as a possible culprit. *Obesity Reviews*, 21(9):e13084, 2020.

[13] Francesco Bonchi, Francesco Gullo, Andreas Kaltenbrunner, and Yana Volkovich. Core decomposition of uncertain graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1316–1325, 2014.

[14] Joaquim Bosch-Barrera, Begoña Martin-Castillo, Maria Buxó, Joan Brunet, José Antonio Encinar, and Javier A Menendez. Silibinin and SARS-CoV-2: Dual Targeting of Host Cytokine Storm and Virus Replication Machinery for Clinical

Management of COVID-19 Patients. *Journal of Clinical Medicine*, 9(6):1770, 2020.

[15] Mehdi Bouhaddou, Danish Memon, Bjoern Meyer, Kris M White, Veronica V Rezelj, Miguel Correa Marrero, Benjamin J Polacco, James E Melnyk, Svenja Ulferts, Robyn M Kaake, et al. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell*, 182(3):685–712, 2020.

[16] Jean Bousquet, Jean-Paul Cristol, Wienczyslawa Czarlewski, Josep M Anto, Adrian Martineau, Tari Haahtela, Susana C Fonseca, Guido Iaccarino, Hubert Blain, Alessandro Fiocchi, et al. Nrf2-interacting nutrients and COVID-19: time for research to develop adaptation strategies. *Clinical and translational allergy*, 10(1):58, 2020.

[17] Grazia Caci, Adriana Albini, Mario Malerba, Douglas M Noonan, Patrizia Pochetti, and Riccardo Polosa. COVID-19 and obesity: dangerous liaisons. *Journal of clinical medicine*, 9(8):2511, 2020.

[18] Valeria Cagno, Gaelle Magliocco, Caroline Tapparel, and Youssef Daali. The tyrosine kinase inhibitor nilotinib inhibits SARS-CoV-2 in vitro. *Basic & Clinical Pharmacology \& Toxicology*, 128(4):621–624, 2021.

[19] Fabrizio Cantini, Laura Niccoli, Carlotta Nannini, Daniela Matarrese, Massimo Edoardo Di Natale, Pamela Lotti, Donatella Aquilini, Giancarlo Landini, Barbara Cimolato, Massimo Antonio Di Pietro, et al. Beneficial impact of Baricitinib in COVID-19 moderate pneumonia; multicentre study. *Journal of Infection*, 81(4):647–679, 2020.

[20] Yang Cao, Jia Wei, Liang Zou, Tiebin Jiang, Gaoxiang Wang, Liting Chen, Liang Huang, Fankai Meng, Lifang Huang, Na Wang, et al. Ruxolitinib in treatment of severe coronavirus disease 2019 (COVID-19): A multicenter, single-blind, randomized controlled trial. *Journal of Allergy and Clinical Immunology*, 146(1):137–146, 2020.

[21] Marco Cascella, Michael Rajnik, Arturo Cuomo, Scott C Dulebohn, and Raffaela Di Napoli. Features, evaluation and treatment coronavirus (COVID-19). *StatPearls: Treasure Island*, 2020.

[22] Anshuman Chandra, Vaishali Gurjar, Imteyaz Qamar, and Nagendra Singh. Identification of potential inhibitors of SARS-COV-2 endoribonuclease (EndoU) from FDA approved drugs: a drug repurposing approach to find therapeutics for COVID-19. *Journal of Biomolecular Structure and Dynamics*, pages 1–11, 2020.

[23] Sai Chen, Jing Zhou, Xiaoqi Ou, Wei Cheng, Yun Qin, Yingqiang Guo, and Yunhan Jiang. Alimentary System is Directly Attacked by SARS-COV-2 and Further Prevents Immune Dysregulation Caused by COVID-19. *International Journal of Clinical Practice*, 75(4):e13893, 2020.

[24] Yuting Cheng, Fang Sun, Luyao Wang, Minjun Gao, Youli Xie, Yu Sun, Huan Liu, Yufeng Yuan, Wei Yi, Zan Huang, et al. Virus-induced p38 MAPK activation facilitates viral infection. *Theranostics*, 10(26):12223, 2020.

[25] UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic acids research*, 38(suppl_1):D142–D148, 2010.

[26] P Conti, G Ronconi, AL Caraffa, CE Gallenga, R Ross, I Frydas, and SK Kritas. Induction of pro-inflammatory cytokines (IL-1 and IL-6) and lung inflammation by Coronavirus-19 (COVI-19 or SARS-CoV-2): anti-inflammatory strategies. *J Biol Regul Homeost Agents*, 34(2):1, 2020.

[27] Domenico Cucinotta and Maurizio Vanelli. Who declares covid-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, 91(1):157, 2020.

[28] Simon Dirmeier, Christopher Dächert, Martijn van Hemert, Ali Tas, Natacha S Ogando, Frank van Kuppeveld, Ralf Bartenschlager, Lars Kaderali, Marco Binder, and Niko Beerenwinkel. Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. *PLoS computational biology*, 16(2):e1007587, 2020.

[29] Kourosh H Ebrahimi, Javier Gilbert-Jaramillo, William S James, and James SO McCullagh. Interferon-stimulated gene products as regulators of central carbon metabolism. *The FEBS journal*, 288(12):3715–3726, 2021.

[30] Mohammed Noor Embi, Nagesswary Ganesan, and Hasidah Mohd Sidek. Is GSK3β a molecular target of chloroquine treatment against COVID-19? *Drug Discoveries & Therapeutics*, 14(2):107–108, 2020.

[31] Lauri Eronen and Hannu Toivonen. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC bioinformatics*, 13(1):119, 2012.

[32] Fatemeh Esfahani, Venkatesh Srinivasan, Alex Thomo, and Kui Wu. Efficient computation of probabilistic core decomposition at web-scale. In *Proceedings of the 22nd International Conference on Extending Database Technology (EDBT)*, pages 325–336, 2019.

[33] Sara Galimberti, Chiara Baldini, Claudia Baratè, Federica Ricci, Serena Balducci, Susanna Grassi, Francesco Ferro, Gabriele Buda, Edoardo Benedetti, Rita Fazzi, et al. The CoV-2 outbreak: how hematologists could help to fight Covid-19. *Pharmacological Research*, 157:104866, 2020.

[34] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):459–468, 2020.

[35] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2):271–280, 2020.

[36] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[37] Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, et al. InterPro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1):D211–D215, 2009.

[38] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(suppl_1):D412–D416, 2009.

[39] Howard M Johnson, Alfred S Lewin, and Chulbul M Ahmed. SOCS, Intrinsic Virulence Factors, and Treatment of COVID-19. *Frontiers in immunology*, 11:2803, 2020.

[40] Yuan-Lin Kang, Yi-ying Chou, Paul W Rothlauf, Zhuoming Liu, Timothy K Soh, David Cureton, James Brett Case, Rita E Chen, Michael S Diamond, Sean PJ Whelan, et al. Inhibition of PIKfyve kinase prevents infection by Zaire ebolavirus and SARS-CoV-2. *Proceedings of the National Academy of Sciences*, 117(34):20803–20813, 2020.

[41] Basil S Karam, Rachel S Morris, Carolyn T Bramante, Michael Puskarich, Emily J Zolfaghari, Sahar Lotfi-Emran, Nicholas E Ingraham, Anthony Charles, David J Odde, and Christopher J Tignanelli. mTOR inhibition in COVID-19: A commentary and review of efficacy in RNA viruses. *Journal of medical virology*, 93(4):1843–1846, 2020.

[42] Rami Bou Khalil. Lithium chloride combination with rapamycin for the treatment of COVID-19 pneumonia. *Medical hypotheses*, 142:109798, 2020.

[43] Wissam Khaouid, Marina Barsky, Venkatesh Srinivasan, and Alex Thomo. K-core decomposition of large networks on a single PC. *Proceedings of the VLDB Endowment*, 9(1):13–23, 2015.

[44] Nilesh Kumar, Bharat Mishra, Adeel Mehmood, Mohammad Athar, and M Shahid Mukhtar. Integrative network biology framework elucidates molecular mechanisms of SARS-CoV-2 pathogenesis. *Iscience*, 23(9):101526, 2020.

[45] Adam Yuh Lin, Michael J Cuttica, Michael G Ison, and Leo I Gordon. Ibrutinib for chronic lymphocytic leukemia in the setting of respiratory failure from severe COVID-19 infection: Case report and literature review. *Ejhaem*, 1(2):596–600, 2020.

[46] Tengwen Liu, Yuhong Guo, Jingxia Zhao, Shasha He, Yunjing Bai, Ning Wang, Yan Lin, Qingquan Liu, and Xiaolong Xu. Systems Pharmacology and Verification of ShenFuHuang Formula in Zebrafish Model Reveal Multi-Scale Treatment Strategy for Septic Syndrome in COVID-19. *Frontiers in pharmacology*, 11:1464, 2020.

[47] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

[48] Krishna P Maremanda, Isaac K Sundar, Dongmei Li, and Irfan Rahman. Age-dependent assessment of genes involved in cellular senescence, telomere, and mitochondrial pathways in human lung tissue of smokers, COPD, and IPF: Associations With SARS-CoV-2 COVID-19 ACE2-TMPRSS2-Furin-DPP4 Axis. *Frontiers in pharmacology*, 11:1356, 2020.

[49] Olanrewaju B Morenikeji, Madeleine Wallace, Ellis Strutton, Kahleel Bernard, Elaine Yip, and Bolaji N Thomas. Integrative network analysis of predicted miRNA-targets regulating expression of immune response genes in bovine coronavirus infection. *Frontiers in genetics*, 11:584392, 2020.

[50] Jingfang Mu, Jiuyue Xu, Leike Zhang, Ting Shu, Di Wu, Muhan Huang, Yujie Ren, Xufang Li, Qing Geng, Yi Xu, et al. Sars-cov-2-encoded nucleocapsid protein acts as a viral suppressor of rna interference in cells. *Science China Life Sciences*, 63(9):1413–1416, 2020.

[51] Jan K Nowak and Jarosław Walkowiak. Lithium and coronaviral infections. A scoping review. *F1000Research*, 9:93, 2020.

[52] Xiuyuan Ou, Yan Liu, Xiaobo Lei, Pei Li, Dan Mi, Lili Ren, Li Guo, Ruixuan Guo, Ting Chen, Jiaxin Hu, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature communications*, 11(1):1620, 2020.

[53] Vid Podpečan, Živa Ramšak, Kristina Gruden, Hannu Toivonen, and Nada Lavrač. Interactive exploration of heterogeneous biological networks with Biomine Explorer. *Bioinformatics*, 35(24):5385–5388, 2019.

[54] Ali A Rabaan, Shamsah H Al-Ahmed, Shafiul Haque, Ranjit Sah, Ruchi Tiwari, Yashpal Singh Malik, Kuldeep Dhama, M Iqbal Yatoo, D Katterine Bonilla-Aldana, Alfonso J Rodriguez-Morales, et al. SARS-CoV-2, SARS-CoV, and MERS-COV: a comparative overview. *Le Infezioni in Medicina*, 28(2):174–184, 2020.

[55] M Janaki Ramaiah. mTOR inhibition and p53 activation, microRNAs: The possible therapy against pandemic COVID-19. *Gene Reports*, 20:100765, 2020.

[56] Kaviyarasi Renu, Mohana Devi Subramaniam, Rituraj Chakraborty, Myakala Haritha, Mahalaxmi Iyer, Geetha Bharathi, Siva Kamalakannan, Vellingiri Balachandar, and VG Abilash. The role of Interleukin-4 in COVID-19 associated male infertility–A hypothesis. *Journal of Reproductive Immunology*, 142:103213, 2020.

[57] Laura Riva, Shuofeng Yuan, Xin Yin, Laura Martin-Sancho, Naoko Matsunaga, Sebastian Burgstaller, Lars Pache, Paul De Jesus, Mitchell V Hull, Max Chang, et al. A Large-scale Drug Repositioning Survey for SARS-CoV-2 Antivirals. *bioRxiv*, 2020.

[58] Laura Riva, Shuofeng Yuan, Xin Yin, Laura Martin-Sancho, Naoko Matsunaga, Lars Pache, Sebastian Burgstaller-Muehlbacher, Paul D De Jesus, Peter Teriete, Mitchell V Hull, et al. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature*, 586(7827):113–119, 2020.

[59] Sandro Rosa and Wilson Santos. Clinical trials on drug repositioning for COVID-19 treatment. *Revista Panamericana de Salud Pública*, 44:e40, 2020.

[60] F La Rosée, HC Bremer, I Gehrke, A Kehr, A Hochhaus, S Birndt, M Fellhauer, M Henkes, B Kumle, SG Russo, et al. The Janus kinase 1/2 inhibitor ruxolitinib in COVID-19 with severe systemic hyperinflammation. *Leukemia*, 34(7):1805–1815, 2020.

[61] Tessa S Schoot, Angèle PM Kerckhoffs, Luuk B Hilbrands, and Rob J Van Marum. Immunosuppressive drugs and COVID-19: A review. *Frontiers in pharmacology*, 11:1333, 2020.

[62] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[63] Farhad Seif, Hossein Aazami, Majid Khoshmirsafa, Monireh Kamali, Monireh Mohsenzadegan, Majid Pornour, and Davood Mansouri. JAK inhibition as a new treatment strategy for patients with COVID-19. *International Archives of Allergy and Immunology*, 181(6):467–475, 2020.

[64] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[65] Hongbo Shi, Wenjing Wang, Jiming Yin, Yabo Ouyang, Lijun Pang, Yingmei Feng, Luxin Qiao, Xianghua Guo, Honglin Shi, Ronghua Jin, et al. The inhibition of IL-2/IL-2R gives rise to CD8+ T cell and lymphocyte decrease through JAK1-STAT5 in critical patients with COVID-19 pneumonia. *Cell Death & Disease*, 11(6):429, 2020.

[66] Justin Stebbing, Venkatesh Krishnan, Stephanie de Bono, Silvia Ottaviani, Giacomo Casalini, Peter J Richardson, Vanessa Monteil, Volker M Lauschke, Ali Mirazimi, Sonia Youhanna, et al. Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients. *EMBO Molecular Medicine*, 12(8):e12697, 2020.

[67] Keiko Taniguchi-Ponciano, Eduardo Vadillo, Héctor Mayani, César Raúl Gonzalez-Bonilla, Javier Torres, Abraham Majluf, Guillermo Flores-Padilla, Niels Wacher-Rodarte, Juan Carlos Galan, Eduardo Ferat-Osorio, et al. Increased expression of hypoxia-induced factor 1α mrna and its related genes in myeloid blood cells from critically ill covid-19 patients. *Annals of Medicine*, 53(1):197–207, 2021.

[68] Ritudhwaj Tiwari, Anurag R Mishra, Flora Mikaeloff, Soham Gupta, Ali Mirazimi, Siddappa N Byrareddy, Ujjwal Neogi, and Debasis Nayak. In silico and in vitro studies reveal complement system drives coagulation cascade in SARS-CoV-2 pathogenesis. *Computational and structural biotechnology journal*, 18:3734–3744, 2020.

[69] Alessandro M Vannucchi, Benedetta Sordi, Alessandro Morettini, Carlo Nozzoli, Loredana Poggesi, Filippo Pieralli, Alessandro Bartoloni, Alessandro Atanasio, Filippo Miselli, Chiara Paoli, et al. Compassionate use of JAK1/2 inhibitor ruxolitinib for severe COVID-19: a prospective observational study. *Leukemia*, 35(4):1121–1133, 2020.

[70] George D Vavougios. SARS-CoV-2 dysregulation of PTBP1 and YWHAE/Z gene expression: A primer of neurodegeneration. *Medical hypotheses*, 144:110212, 2020.

[71] Nina Verstraete, Giuseppe Jurman, Giulia Bertagnolli, Arsham Ghavasieh, Vera Pancaldi, and Manlio De Domenico. CovMulNet19, Integrating Proteins, Diseases, Drugs, and Symptoms: A Network Medicine Approach to COVID-19. *Network and systems medicine*, 3(1):130–141, 2020.

[72] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21, 2007.

[73] Zofia Wicik, Ceren Eyileten, Daniel Jakubik, Sérgio N Simões, David C Martins, Rodrigo Pavão, Jolanta M Siller-Matula, and Marek Postula. ACE2 interaction networks in COVID-19: a physiological framework for prediction of outcome in patients with cardiovascular risk factors. *Journal of Clinical Medicine*, 9(11):3743, 2020.

[74] Bryan J Winn. Is there a role for insulin-like growth factor inhibition in the treatment of COVID-19-related adult respiratory distress syndrome? *Medical Hypotheses*, 144:110167, 2020.

[75] Dandan Wu and Xuexian O Yang. TH17 responses in cytokine storm of COVID-19: An emerging target of JAK2 inhibitor Fedratinib. *Journal of Microbiology, Immunology and Infection*, 53(3):368–370, 2020.

[76] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.

[77] Qi-Dong Xia, Yang Xun, Jun-Lin Lu, Yu-Chao Lu, Yuan-Yuan Yang, Peng Zhou, Jia Hu, Cong Li, and Shao-Gang Wang. Network pharmacology and molecular docking analyses on Lianhua Qingwen capsule indicate Akt1 is a potential target to treat and prevent COVID-19. *Cell proliferation*, 53(12):e12949, 2020.

[78] Zi-jian Xie, Joel Novograd, Yaakov Itzkowitz, Ariel Sher, Yosef D Buchen, Komal Sodhi, Nader G Abraham, and Joseph I Shapiro. The pivotal role of adipocyte-na k peptide in reversing systemic inflammation in obesity and covid-19 in the development of heart failure. *Antioxidants*, 9(11):1129, 2020.

[79] Swamy Yeleswaram, Paul Smith, Timothy Burn, Maryanne Covington, Ashish Juvekar, Yanlong Li, Peg Squier, and Peter Langmuir. Inhibition of cytokine signaling by ruxolitinib and implications for COVID-19 treatment. *Clinical Immunology*, 218:108517, 2020.

[80] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[81] Jiancheng Zhang, Bing Xie, and Kenji Hashimoto. Current status of potential therapeutic candidates for the COVID-19 crisis. *Brain, Behavior, and Immunity*, 87:59–73, 2020.

[82] Xiuhong Zhang, Yan Zhang, Weizhen Qiao, Ji Zhang, and Zhigang Qi. Baricitinib, a drug with potential effect to prevent SARS-COV-2 from entering target cells and control cytokine storm induced by COVID-19. *International Immunopharmacology*, 86:106749, 2020.

[83] Jing Zhao, Saisai Tian, Dong Lu, Jian Yang, Huawu Zeng, Feng Zhang, Dongzhu Tu, Guangbo Ge, Yuejuan Zheng, Ting Shi, et al. Systems pharmacological study illustrates the immune regulation, anti-infection, anti-inflammation, and multi-organ protection mechanism of Qing-Fei-Pai-Du decoction in the treatment of COVID-19. *Phytomedicine*, 85:153315, 2021.

[84] Wen-jiang Zheng, Qian Yan, Yong-shi Ni, Shao-feng Zhan, Liu-liu Yang, Hong-fa Zhuang, Xiao-hong Liu, and Yong Jiang. Examining the effector mechanisms of Xuebijing injection on COVID-19 based on network pharmacology. *BioData mining*, 13:17, 2020.

[85] Yunfeng Zheng, Renfeng Li, and Shunai Liu. Immunoregulation with mTOR inhibitors to prevent COVID-19 severity: A novel intervention strategy beyond vaccines and specific antiviral medicines. *Journal of Medical Virology*, 92(9):1495–1500, 2020.

[86] Yadi Zhou, Yuan Hou, Jiayu Shen, Reena Mehra, Asha Kallianpur, Daniel A Culver, Michaela U Gack, Samar Farha, Joe Zein, Suzy Comhair, et al. A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19. *PLoS biology*, 18(11):e3000970, 2020.

[87] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications*, 10(1):1523, 2019.

[88] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in China, 2019. *New England journal of medicine*, 382(8):727–733, 2020.

[89] Pengpeng Zhu, Chenfei Lv, Chengxiu Fang, Xing Peng, Hao Sheng, Peng Xiao, Nishant Kumar Ojha, Yan Yan, Min Liao, and Jiyong Zhou. Heat shock protein member 8 is an attachment factor for infectious Bronchitis virus. *Frontiers in microbiology*, 11:1630, 2020.

[90] Zhenjie Zhuang, Xiaoying Zhong, Huanhuan Zhang, Huiqi Chen, Boxiang Huang, Dongqun Lin, and Junmao Wen. Exploring the Potential Mechanism of Shufeng Jiedu Capsule for Treating COVID-19 by Comprehensive Network Pharmacological Approaches and Molecular Docking Validation. *Combinatorial chemistry & high throughput screening*, 2020.