**Sapienza University of Rome**

-

M.Sc. in Artificial Intelligence and Robotics

M.Sc. Thesis

# Network Connectivity Analysis for SARS-CoV-2-Human Protein-Protein Interaction

Thesis Advisor
**Prof. Francesca Cuomo**

Thesis Coadvisors
**Prof. Stefania Colonnese**
**Prof. Manuela Petti**

Candidate
**Giulia Cassara'**
**1856973**

**Academic Year 2019-2020**

*Everything is connected...*

# Abstract

Is it possible to find an effective drug treatment by simply analyzing a graph? In recent years, a new discipline emerged, applying network theory to the identification, prevention, and treatment of diseases: network medicine.

In this thesis work, I present a case study for the search of a drug treatment for COVID-19, using a technique based on the analysis of the gene interaction between the SARS-CoV-2 organism, the human organism and drugs.

The work focused on the extraction of the network motifs and the prioritization of those drugs with higher centrality measures.

This work showed promising results, being able to prioritize drugs that later have been extensively discussed in medical literature for COVID-19 treatment.

# Contents

# Chapter 1

# Introduction

## 1.1 Network biology and network medicine

In the last years, a new biological field has emerged: *network biology.* Opposed to a reductionist approach, in which only single cells are studied and observed, the approach of network biology is holistic, using network theory to understand the structure and dynamics of the complex intercellular web of interactions, that contribute to the structure and function of a living cell [1].

It comes natural to also extend this kind of considerations to diseases, as a disease is rarely a consequence of an abnormality of a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. Network medicine is a new interdisciplinary field using network topology and network dynamics towards identifying diseases and developing medical drugs [2].

*Network medicine* is based on a series of hypotheses that link network structure to biological functions and diseases [2].
Some hypotheses tells us the way biological entities like proteins interact in the

interconnected disease network. Mutations in interacting proteins often lead to similar disease phenotypes.

It is necessary to distinct between different type of "modules". A **topological module** represents a locally-dense neighbourhood in a network, such that nodes have a higher tendency to link to nodes within the same local neighbourhood than to nodes outside it. Such modules can be identified using network clustering algorithms that are unaware of the function of individual nodes.
A **functional module** represents the aggregation of nodes of similar or related function in the same network neighbourhood.
Finally, a **disease module** represents a group of network components that together contribute to a cellular function. By disrupting one of those components, we obtain a particular disease phenotype.

In network biology, we assume that those three modules overlaps: cellular components that form a topological module also correspond to a functional module, and a disease is a result of the breakdown of a particular functional module, intimating that a functional module is also a disease module. It's important to note that a disease module is defined in relation to a particular disease and, accordingly, each disease has its own unique disease module. Moreover, a gene, a protein or a metabolite can be implicated in several disease modules, which means that different disease modules can overlap. Cellular components associated with a specific disease phenotype show a tendency to cluster in the same network neighbourhood.

Another important assumption is that all cellular components that belong to the same topological, functional or disease module have a high likelihood of being involved in the same disease.

## 1.2 Drug Repurposing

*Drug repurposing*, or repositioning, is a technique whereby existing drugs are used to treat emerging and challenging diseases, including novel coronavirus disease 2019 (COVID-19), but also cancer and other rare diseases.
Drug repurposing has become a promising approach because of the opportunity for reduced development timelines and overall costs compared to classical Drug Discovery. In fact, drug discovery processes are getting slower and more difficult over time; this observation was formalized with the concept of "Eroom's Law" [1] [3].

As the orignal Moore's Law, it's based on the observation that, with the huge advancement of technology, the number of new drugs approved per billion US dollars spent on R&D has halved roughly every 9 years since 1950. This leads, for companies that develop new drugs, to higher time-to-market and exponentially-increasing costs.

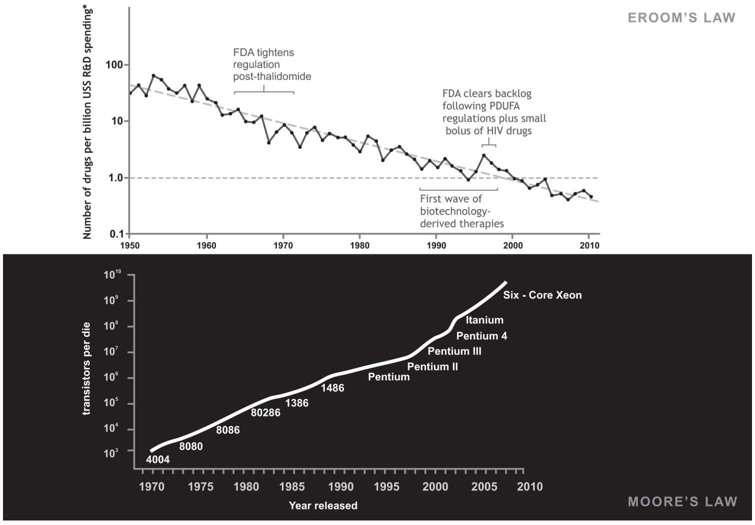


**Figure 1.1:** Eroom's Law in drug development

[1] Which is the "Moore's Law" spelled backward.

The main advantages of drug repositioning are:

- Faster drug discovery at a fraction of costs due to the reduced steps of clinical trials.

- Existing pharmaceutical supply chains could facilitate the distribution of the drug

- By lowering the effort on the development of new drugs, researches can focus on the possibility of combining with other drugs for a more effective treatment.

Successful examples of drug repositioning include the use of *sildenafil* for erectile dysfunction and pulmonary hypertension. Pfizer was seeking a drug for angina when it originally created *sildenafil*, commercially known as Viagra, in the 1980s. The desired cardiovascular effects were not observed on the healthy volunteers. However, on of them reported in their questionnaires that they had had unusually strong and persistent erections [4] .

*Thalidomide*, which originally was developed in the 1950s as a sedative for pregnant women (morning sickness), was subsequently used as a therapy for leprosy, and approved by the FDA in 2006 for treatment of multiple myeloma [5].

Combining several existing antivirals and the knowledge gained from past SARS and MERS outbreaks, HIV/AIDS, and malaria, have been researched as potential COVID-19 treatments, like *Remdesivir* and *Chloroquine* [6]. However, most of the successful examples of drug repositioning come from clinical observations, like the case of Viagra [4]. More often, the underlying molecular mechanisms are not clear and a general awareness is that many targets have not yet been discovered. Therefore, computational methods are expected to effectively reposition drugs against various diseases. Some existing drugs may have a potential positive therapeutic effect with new targets.

Drug repositioning is grounded on two key observations:

1. Complex diseases are generally caused by the collective abnormalities of a number of correlated genes
2. Biological processes will be perturbed due to the manifestation of the drug's effect

These two facts are also the basis of the hypotheses of network biology. The two fields, network biology and drug repositioning, aims to find newly emerged properties at a network level, and to investigate how intra-cellular and extra-cellular systems induce different biological pathophenotypes under different conditions.

## 1.3 Thesis Goal

Though there are many tools and researches on network-based approaches for drug repurposing, very few of them rely entirely on network theory. At the time of the writing of this thesis, very little effort was spent by the scientific community in analyzing the network topology and connectivity of the protein-protein interactions between SARS-CoV-2 proteins and human proteins from a purely computational perspective.

I developed a framework that, from the protein-protein interactions between SARS-CoV-2 proteins and human proteins and the drug-target interactions, is able to prioritize effective drugs that form a specific network motif within the artificially-built interactome.

I've also explored the topological and connectivity properties of the interactome by performing several centrality measures and the spectrum of the graph's Laplacian. I have also investigated how the centrality and the spectrum changes according to different networks from different sources.

# Chapter 2

# Graph Theory

## 2.1 Definitions and mathematical preliminary

### 2.1.1 Graph definitions

**Undirected single graph**

A graph $G$ can be defined as a pair of $(V, E)$, where $V$ is a set of vertices representing the nodes and $E$ is a set of edges representing the connection between nodes. Two nodes $i$ and $j$ are connected if and only if $i, j \in V$. In this case, we say that $i$ and $j$ are neighbors. Protein-Protein Interactions (PPI) are best representented with undirected graph.

**Directed graph**

If we instead consider not only the interaction, but the directionality of the interaction, we say that we have a directed graph. Formally, a directed graph is defined as an ordered triple $G = (V, E, f)$, where $f$ is a function that maps each element in $E$ to an ordered pair of vertices in $V$; so an edge $E = (i, j)$ is considered to have direction from $i$ to $j$. Directed graphs are suitable for the representation of schemas describing biological pathways or procedures which show the sequential interaction of elements at one or multiple time points and the flow of information throughout the network.

### Adjacency matrix

Given a graph $G = (V, E)$, the adjacency matrix representation[1] is a matrix with rows and columns labeled by graph vertices, with a 1 or a 0 in position $(i, j)$ according to whether $i$ and $j$ are adjacent or not. For a simple graph with no self-loops, the adjacency matrix must have 0s on the diagonal. For an undirected graph, the adjacency matrix is symmetric, so only upper or lower triangular part of the matrix is necessary, which halves of the memory to be allocated for the matrix.

### Subgraph

A graph $G' = (V', E')$ is a subgraph of $G$, if $V' \subseteq V$ and $E' \in E$. If $G' = (V', E')$ is a subgraph of $G$ and $E'$ contains all edges $e_{uv} \in E$ such that $u, v \in V'$, then $G'$ is called an induced subgraph of $G$.

### Isomorphism

Two graphs, $G' = (V', E')$ and $G'' = (V'', E'')$, are called isomorphic if there exists a bijection $h : V' \Rightarrow V''$ such that any 2 nodes $u$ and $v$ of $G'$ are adjacent in $G'$ if and only if $h(u)$ and $h(v)$ are adjacent in $G''$.

[1] Sometimes also called the connection matrix.

## 2.2 Network properties

### 2.2.1 Graph density

The graph density shows how sparse or dense a graph is according to the number of connections per node set.
For an undirected graph, the density is:

$$\frac{2|E|}{|V|(|V|-1)}$$

For a directed graph, the density is:

$$\frac{|E|}{|V|(|V|-1)}$$

Typically, connectivity densities of biological gene networks are lower than 0.1, so they are sparsely connected. Quoting the authors of the paper [7]:

> "*This indicates that sparse networks are actually more robust if the costs of complexity are accounted for. If true, then evolution should seek to optimize the costs and benefits of complexity with a parsimonious network structure, a network topology that is sparsely connected and not unnecessarily complex, by seeking an optimal topological ensemble of interactions that best meets the network's functional requirements under its normal range of operating conditions*"

## 2.3 Network Motifs

Network Motifs represent patterns in complex networks occurring significantly more often than in randomized networks [8].

Motif determination can give information about the properties and the characteristics of a network. Some motifs have been found to be associated with optimized biological functions, like in the case of positive and negative feedback loops, oscillators or bifans [2]. Formally, a motif $G' = (V', E')$ is an $n$-node subgraph of $G$, where $|V'|$ is $n$.

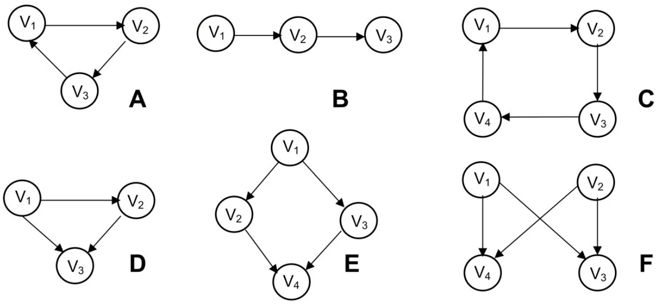


**Figure 2.1:** Network Motifs. A) Feed-forward loop. Type of networks: protein, neuron, electronic. B) Three chain. Type of network: food webs. C) Four node feedback. Type of network: gene regulatory, electronic. D) Three node feedback. Type of network: gene regulatory, electronic. E) Bi-parallel. Type of network: gene regulatory, biochemical. F) Bi-Fan.

## 2.4 Network centralities

Once the biological network is modeled as a graph, centrality analysis is particularly useful for understanding the underlying biological processes. Based on centrality measures the graph elements such as vertices and edges can be ranked from different points of view - *"What are the most connected proteins?"*, *"What protein is acting as a bridge?"*, *"What protein is acting as a super-regulator in the disease?"*. Based on the network hypotheses, top ranked elements in the graph are supposed to play an important role in the network.

### 2.4.1 Degree centralities

Degree Centrality shows that an important node is involved in a large number of interactions.

For an undirected graph, the node degree is the number of links in that node. For directed graphs, each node is obviously characterized by two degree centralities: the *in-degree* and the *out-degree.* Hub is a node with very high degree. Scale-free networks tend to contain hubs. The removal of such central nodes has great impact on the topology of the network. It has been shown that biological networks tend to be robust against random perturbations, but disruption of hubs often leads to system failure [9]. Not surprisingly, in this recent pre-print work, a new strategy for designing combination therapies is proposed by targeting hubs in cancer networks that are not associated with relevant toxicity networks [10]

### 2.4.2 Betweenness Centrality

Given two distinct nodes $i, j, w \in V(G)$, let $\sigma_{ij}$ be the total number of shortest paths between $i$ and $j$ and $\sigma_{ij}(w)$ be the number of shortest paths from $i$ to $j$ that pass through $w$. Moreover, for $w \in V(G)$, let $V(i)$ denote the set of all ordered pairs, $(i, j)$ in $V(G) \times V(G)$ such that $i, j, w$ are all distinct. Then, the Betweenness Centrality is calculated as:

$$c_b(w) = \sum_{(i,f)\in V(w)} \frac{\sigma_{if}(w)}{\sigma_{if}}$$

Nodes which are intermediary between neighbors rank higher in betweenness centrality. Without these nodes, there would be less efficient or no way for two neighbors to communicate with each other. Proteins with high betweenness are “bottlenecks”, they have a key role as connectors with essential functional and dynamic properties. The betweenness of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating proteins. In signaling modules, proteins with high betweenness are likely crucial to maintain functionality and coherence of signaling mechanisms, thus, they may form good targets for drug discovery [11].

### 2.4.3 Closeness Centrality

A central node is one that is close, on average, to other nodes.
Closeness indicates important nodes that can communicate quickly with other nodes of the network. It is a measure of how fast information spreads from a given node to other reachable nodes in the network.
Let $G = (V, E)$ be an undirected graph. Then, the closeness centrality is defined as:

$$c_{clo}(i) = \frac{1}{\sum_{t \in V} distance(i, j)}$$

The denominator is the distance or the shortest path $p$ between the nodes $i$ and $j$.
Closeness takes high values for vertices that are separated from all others by only a short geodesic distance on average.

High-closeness nodes might have:

- better access to information in other vertices or
- more direct influence on other vertices

Closeness centrality has been used to identify the top central metabolites in genome-based large-scale metabolic networks [12].

### 2.4.4 Clustering Coefficient

Clustering coefficient, generally speaking, is the measurement that shows the tendency of a graph to be divided into clusters.

High clustering coefficient means that the network contains communities or groups of nodes that are densely connected internally. In biological networks, finding these communities is very important, because they can reflect functional modules and protein complexes.

Following an analogy from the social sciences, *"the friends of my friends are my friends"*.
We want to see, for a given vertex $v$, to what extent the neighbors of $v$ are also neighbors of each other. In other words, to what extent are vertices adjacent to $v$ also adjacent to each other.

### Local Clustering Coefficient

In undirected networks, the clustering coefficient of a node n is defined as:

$$C_n = \frac{2e_n}{k_n(k_{n-1})}$$

- $k_n$ is the number of neighbors of n
- $e_n$ is the number of connected pairs between all neighbors of n

### Global Clustering Coefficient

The global (or average) Clustering Coefficient is defined as:

$$C_{avg} = \frac{1}{N}\Sigma_i^N \frac{E_i}{k_i(k_{i-1})}$$

The closer the local clustering coefficient is to 1, the more likely it is for the network to form clusters.

### 2.4.5 Eigenvector Centrality

Eigenvector centrality is a basic extension of degree centrality, which defines centrality of a node as proportional to its neighbors' importance, not on its number of connections. The centrality of each vertex is proportional to the sum of the centralities of its neighbors (recursive definition).
Let $G = (V, E)$ be an undirected graph and $A$ the adjacency matrix of network $G$.
The eigenvector centrality is the eigenvector $C_{eiv}$ of the largest eigenvalue $\lambda_{max}$ in absolute value such that $\lambda Ceiv = ACeiv$.

A protein with a very high Eigenvector is a protein interacting with several important proteins (regulating them or being regulated by them), thus suggesting a central super-regulatory role or a critical target of a regulatory pathway.

Pharmacological modulation (for instance inhibition) of a high Eigenvector protein may generate broad biological effects but potentially characterized by consistent side-effects.

## 2.5 Laplacian Matrix

Given a graph $G = (V, E)$, The Laplacian matrix of a graph G is defined as $L = D - A$, where D, called the degree matrix, is a diagonal matrix with the $j - th$ diagonal element $d_{jj} = \sum_{i=1} aij$ and $A$ is the adjacency matrix of $G$. $L(G)$ is positive semidefinite and singular matrix [13]

## 2.6 Three basic network models

When analyzing a graph, it is a good strategy to compare the graph structure and topology with three different basic network models that are used as reference in network theory.
These three models are the Erdős–Rényi's random graph model, Watts–Strogatz's small–world model, and Barabási–Albert's scale–free model.

### 2.6.1 Erdős–Rényi random graph

Erdős and Rényi proposed a very simple model [14] for a random graph. For a graph constructed from $n$ vertices, the existence of a connection between each vertex pair depends on a fixed probability $p$.
That means that while constructing the network, one could make a decision to connect (respectively not connect) a pair of vertices by an edge with the probability $p$ (or $1-p$). If we extend the reasoning, the probability of obtaining any one particular random graph with $m$ edges is $p^m(1-p)^(n-m)$.
Many different properties emerges with different values of p; when p is small, the graph appears almost disconnected. For $p > \frac{1}{n}$ the graph tend to be connected.

The main features of Erdős–Rényi graphs:

- short average path lenghts.
- low clustering coefficient.
- low probability to have hubs.

### 2.6.2 Watts–Strogatz's small-world network

Inspired by the small world phenomenon described by Milgram in 1967, Watts and Strogatz introduced a network model with high clustering coefficient and

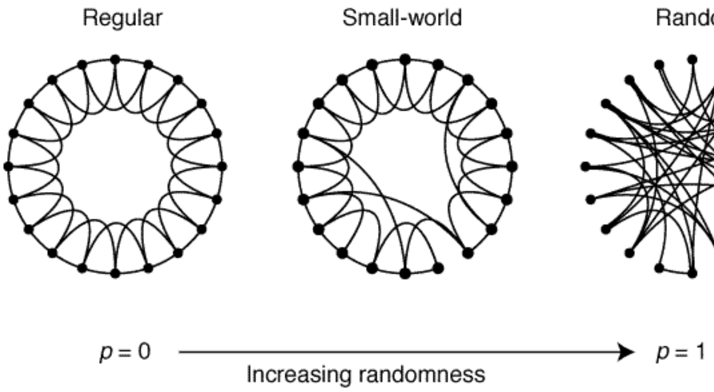


**Figure 2.2:** Random rewiring procedure for interpolating between a regular ring lattice and a random network, without altering the number of vertices or edges in the graph.

a low average path length [15].

Starting from a ring lattice with n vertices and k edges per vertex, each edge is rewired at random with probability $p$. This construction allows us to 'tune' the graph between regularity ($p = 0$) and disorder ($p = 1$), and thereby to probe the intermediate region $0 < p < 1$, about which little is known. Basically, this model is an interpolation between regular and random networks as shown in the figure.

The main features of a Watts–Strogatz's graph:

- short average path lenght.
- high clustering coefficient.
- Very peaked degree distribution.

### 2.6.3 Barabási–Albert's scale–free network

Barabási and Albert proposed a growing model with preferential attachment of new vertex ("rich gets richer", the more connected a node is, the more likely it is to receive new links). A graph constructed from this model follows a power law degree distribution [16], with $P(k) \sim k^( - \gamma)$. In contrast with Erdős–Rényi and Watts–Strogatz's network models, the power-law tail characterizing P(k)

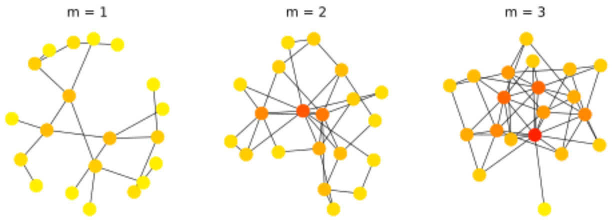


**Figure 2.3:** Display of three graphs generated with the Barabasi-Albert (BA) model. Each has 20 nodes and a parameter of attachment m as specified

for BA networks indicates that highly connected (large k) nodes have a larger chance of occurring, dominating the connectivity.

The main features of a Barabási–Albert graph:

- shorter average path length than a random graph.
- clustering coefficient follows a power law with the node degree.
- Scale-free degree distribution.

# Chapter 3

# Methods

## 3.1 Software used

The steps of data collection, processing and encoding, was written in Python. Python is an interpreted, high-level, general-purpose programming language [17]. The choice of Python was due to its simple syntax and easy accessibility to libraries and packages for data manipulation.
The modules used are:

- Python 3.8.5
- Pandas
- csv
- BeautifulSoup
- json
- numpy

The network centrality measures, spectral analysis and visualization steps were written in Matlab.In fact, Matlab provides effective matrix manipulation and plotting.
I've used the GraphLab [18] toolbox for importing graph from `.txt` data and plot relevant network measurements and distributions.

## 3.2 Data Collection

### 3.2.1 Protein - Protein Interactions

Proteins rarely act alone. Often, they team up and have intricate physicochemical dynamic connections to undertake biological functions at both cellular and systems levels. A critical step towards unraveling the complex molecular relationships in living systems is the mapping of protein-to-protein physical "interactions". The complete map of protein interactions that can occur in a living organism is called the **interactome** [19]. Proteins are nodes, linked by their interaction. The nodes and links together form a network, or, in more formal mathematical language, a **graph**.

In general, not all the Protein-Protein-Interactions (PPI) are equal. There is a strong need to distinguish between "experimental" PPIs and "predicted" PPIs in order to avoid misinterpretation of the results provided by one or the other approach. Both types of data can be useful, but it is not the same to test an interaction between protein A and B by Yeast two-hybrid as it is to infer a possible interaction between protein A and B based on their gene co-expression profile. In the first situation, the PPI is experimentally proven, while in the second the PPI is predicted from experimental data obtained for the corresponding genes, which does not prove a direct protein interaction.

A *primary* PPI database extracts PPIs from the experimental evidence reported in the peer-reviewed scientific publications. An example of primary database are: BioGRID [20], IntAct2 [21] and MatrixDB [22].

On the other hand, *predictive* databases like STRING [23] combine the experimentally inferred data taken from primary databases with computational predictions of molecular interactions. They are be extremely useful when very few data is available.

### 3.2.2 Virus - Host Interactions

Virus–host interactions represents the viral and host processes that occur during viral infection, which enable both organisms to respond on another. One of the key questions in virology is how viruses, even though they encode relatively few genes, are able to gain temporary or constant control over their hosts. To understand the *pathogenicity* of a virus, it is important to gain knowledge on the function of the individual viral proteins in the host cell, and on the consequences of these interactions on cellular signaling pathways. Viral and human interactome analyses can help elucidate how different viruses target human proteins.

The discovery of new therapeutic targets is possible by studying the infection maps, where the viral infection can be visualized as perturbation of the human PPI network, and by identifying the biological functions that are impaired by these perturbations

For the tasks of collecting virus-host interactions and protein-protein interactions, the latest version of BioGRID has been used as a primary biological database that stores PPIs, genetic interactions and even chemical interactions [20].

BioGRID was created in 2003 and provides lists of interactions for all major species. Many biological databases has now integrated a special issue regarding information of the Coronavirus.

The BioGRID COVID-19 Coronavirus Project include interactions from 221 publications and preliminary reports. The total number of Coronavirus-related interactions has reached the number of 21,405 and 47 post-translational modifications including 19,524 interactions and 47 post-translational modifications from 127 publications and preliminary reports for SARS-CoV-2 (COVID-19) and 14 coronavirus-related CRISPR screens.

### 3.2.3 Drug - Target Interactions

The drug-target databases contain information and interactions of drugs and drug targets.The analysis of the human target-protein network topological properties[1] has revealed that successful targets are more highly connected to other proteins and have higher network betweenness than other proteins [24].

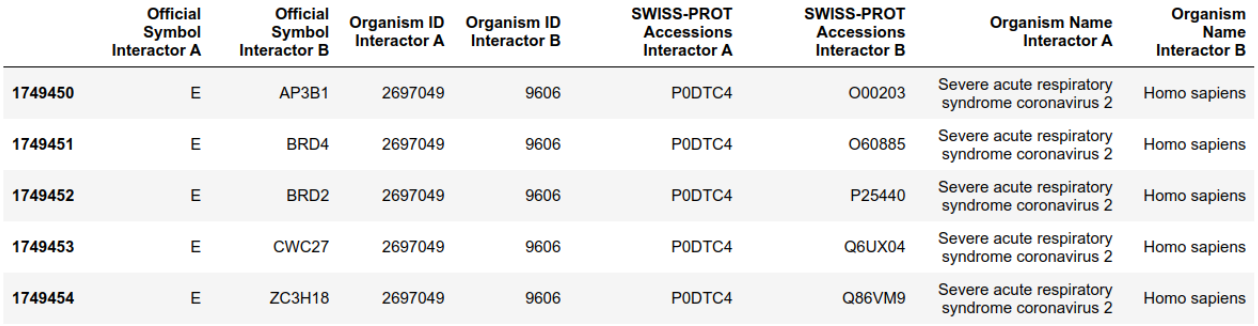
The *DrugBank* database is a richly annotated bioinformatics and cheminformatics resource that combines detailed drug data (e.g. chemical, pharmacological and pharmaceutical) with comprehensive target information (e.g. sequence, structure and pathway). The database is updated frequently. DrugBank Online is widely used and is enabling major advancements across the data-driven medicine industry.

#### Parsing and post-processing

The BioGRID database parsing and post-processing has been done in Python using libraries for data manipulation like *pandas* and *csv*. For the purposes of this project, I've considered only the interactions between Human proteins and Sars-COV-2.
In Fig.3.1 is depicted the *pandas* data frame generated by parsing the BioGRID data.

The column name "Official Symbol Interactor X" represents the official gene

[1] Derived from the human target-protein interaction data



| | Official Symbol Interactor A | Official Symbol Interactor B | Organism ID Interactor A | Organism ID Interactor B | SWISS-PROT Accessions Interactor A | SWISS-PROT Accessions Interactor B | Organism Name Interactor A | Organism Name Interactor B |
|---|---|---|---|---|---|---|---|---|
| **1749450** | E | AP3B1 | 2697049 | 9606 | P0DTC4 | O00203 | Severe acute respiratory syndrome coronavirus 2 | Homo sapiens |
| **1749451** | E | BRD4 | 2697049 | 9606 | P0DTC4 | O60885 | Severe acute respiratory syndrome coronavirus 2 | Homo sapiens |
| **1749452** | E | BRD2 | 2697049 | 9606 | P0DTC4 | P25440 | Severe acute respiratory syndrome coronavirus 2 | Homo sapiens |
| **1749453** | E | CWC27 | 2697049 | 9606 | P0DTC4 | Q6UX04 | Severe acute respiratory syndrome coronavirus 2 | Homo sapiens |
| **1749454** | E | ZC3H18 | 2697049 | 9606 | P0DTC4 | Q86VM9 | Severe acute respiratory syndrome coronavirus 2 | Homo sapiens |

**Figure 3.1:** *pandas* data frame generated by parsing the BioGRID data

symbol approved by HGNC. The “SWISS-PROT Accessions Interactor X” represents the Uniprot name of the gene, which is used to query it to external databases, in this case, DrugBank.

### 3.2.4 Collecting the drug-target interactions

DrugBank database is organized as an `xml` file containing all informations about drug-target.

Starting from the original `xml` file from DrugBank, it was necessary to create an inverted index which contains the association between targets and those drugs whose group is in one of the following lists: “approved”, “experimental”, “nutraceutical”, “investigational”.

The inverted index contains, for every protein, all the drugs that targets it. Then we filter out only those protein that appears in the “Sars-Cov-2-Human” interactome and we reverse this structure into an index:

$$Drug : [List\ of\ protein\ target\ from\ the\ disease\ module]$$

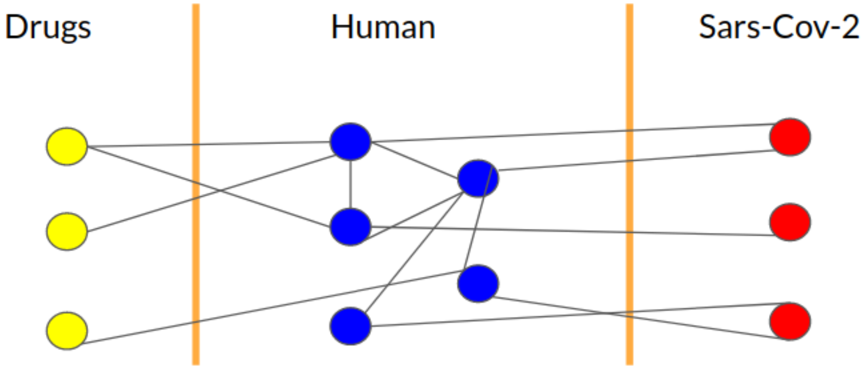
This final data structure is a dictionary containing the intersection between Sars-CoV-2 targets and Drug targets. It is dumped in `json` format.

### 3.2.5 Putting it all together

We have a tri-partite graph with:

- 30 SARS-CoV-2 genes interacting with at least one human gene
- 4355 Human genes interacting with at least one SARS-CoV-2 gene and interacting each other (sars-human + human-human interactions)
- 2002 drugs interacting with at least one human protein (target)

Below, there is a logical graphical representation of the interactions obtained so far.



### 3.2.6 Encoding

All the future analysis and visualization are done in Matlab, so data must be representented in a uniform manner. Every node must be encoded by a natural number in a monotonic sequence in order to be processed as a graph by Matlab.

- Numbers from 0 to 29 are assigned to SARS-CoV-2 genes
- Numbers from 30 to 4384 are assigned Human proteins
- Numbers from 4385 to 6386 are assigned to drugs

## 3.3 Network Analysis & Visualization

### 3.3.1 Centrality Measures - ranking

Now we will explore and visualize the interactome centrality measures by relating them to that obtained from the graph models of Erdos-Renyi, Watts Strogatz and Barabasi-Albert.

The Drug-Virus-Host interactome is a fully connected graph with 6387 nodes and 113737 interactions. The graph is sparse, with the value of the density almost 0.056. The other graph models were build with the same number of nodes and density.

The scale-free network was generated in Matlab by following [16].

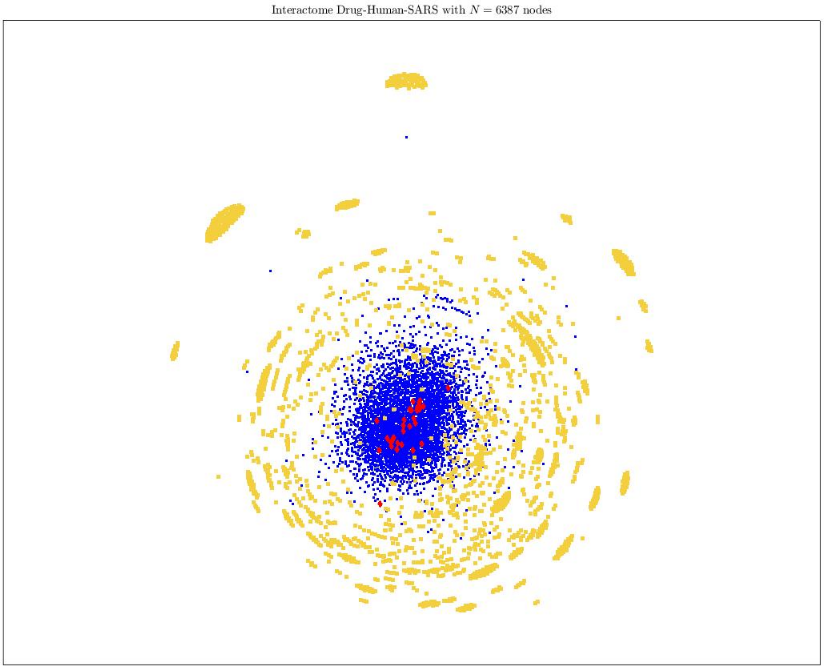


**Figure 3.2:** Drug-Virus-Host interactome. Nodes in yellow are drugs, in blue are human proteins, in red SARS-CoV-2 proteins

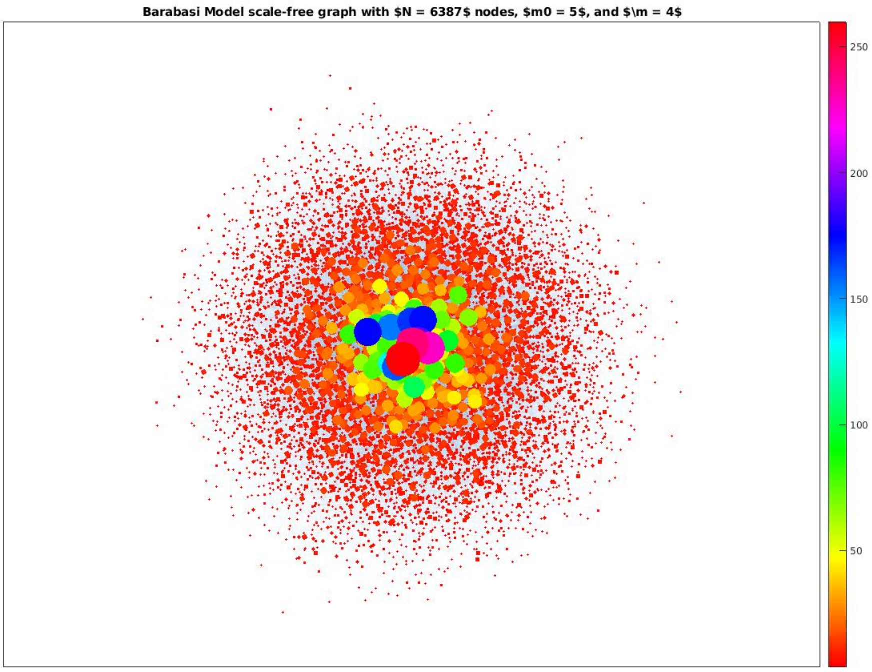


**Figure 3.3:** Barabasi-Albert scale-free network

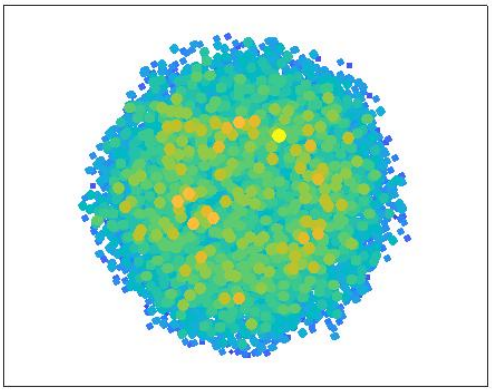


**Figure 3.4:** Watts-Strogatz model

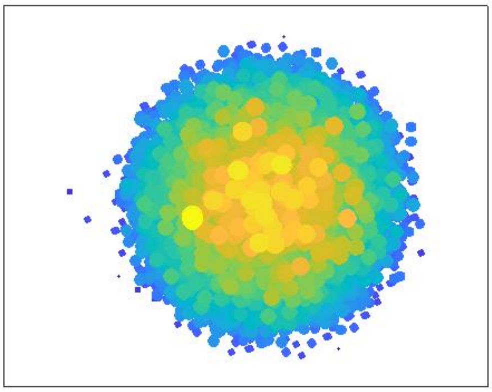
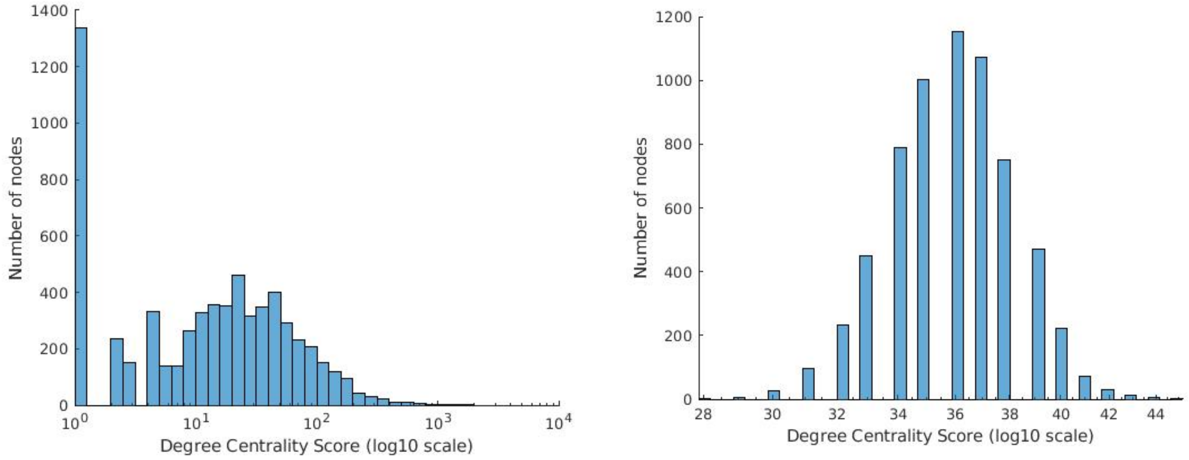


**Figure 3.5:** Erdos-Renyi random graph

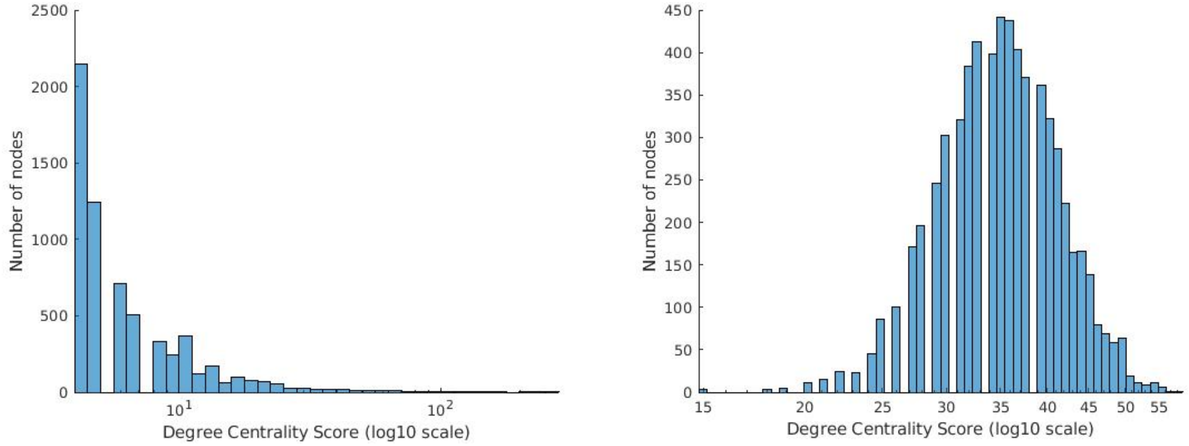
The high dimensionality of the data makes the classical visualization method like heatmap poorly indicative. In the following paragraphs It will be showed an histogram of the centrality measures in logarithmic scale for all the graph models. The X-axis represents the centrality score, the y-axis is the number of nodes that has that value.

The behaviour of the Drug-Virus-Host degree centrality show that isolated or poorly connected nodes constitutes the largest number. This behaviour is similar to that of the Scale-free network 3.3, although in the latter few nodes are the most connected. In the interactome the frequency of the most connected nodes doesn't follow a specific law.



a) b)

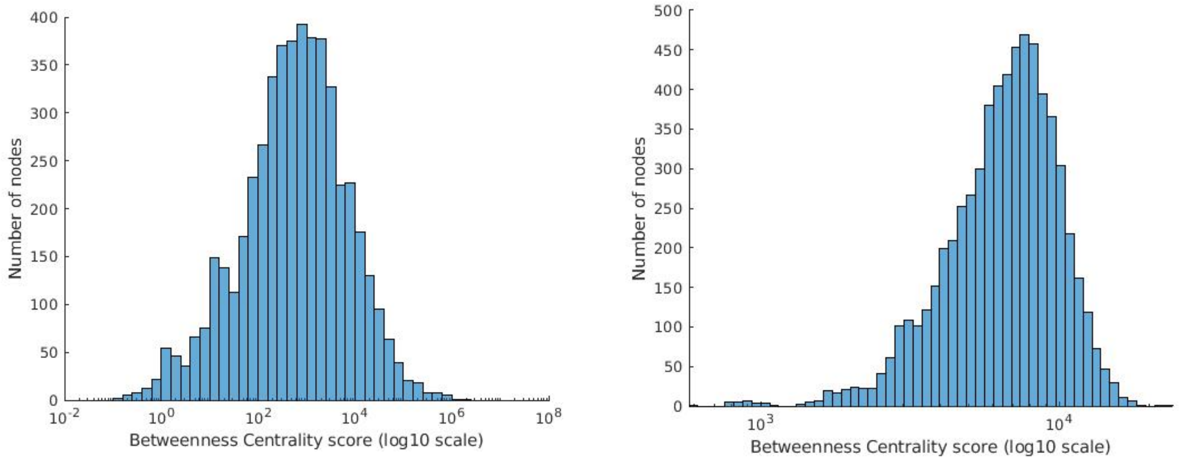
**Figure 3.6:** a) Degree centrality of the Drug-Virus-Host interactome b) Degree centrality of the Watts-Strogatz network model



a) b)

**Figure 3.7:** c) Degree centrality of the Scale-free network d) Degree centrality of the Random graph

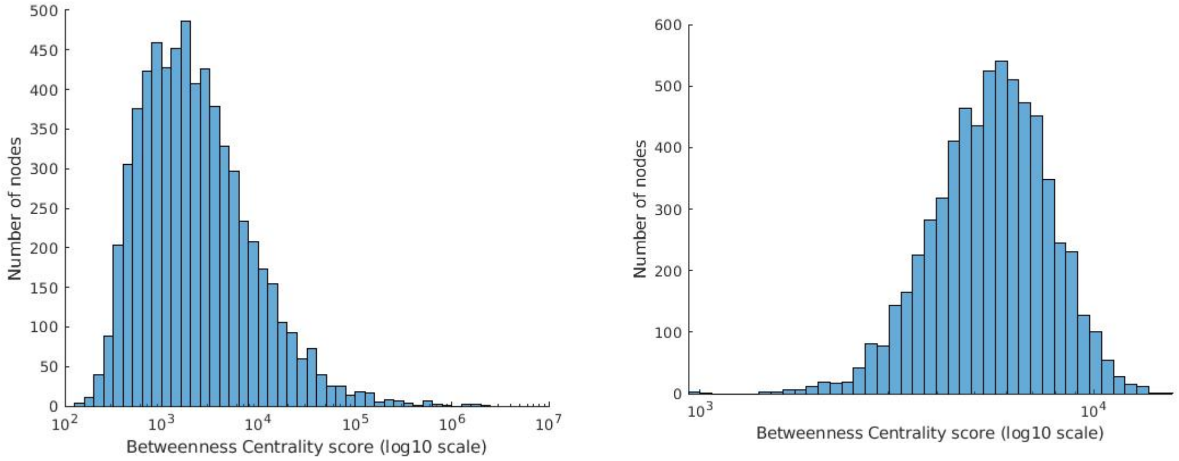
The histogram of the betweenness centrality score of the Drug-Virus-Host interactome follows a normal distribution with the greatest range of values ($10^{10}$) compared to the other graphs.



a)

b)

**Figure 3.8:** a) Betweenness centrality of the Drug-Virus-Host interactome b) Betweenness centrality of Watts-Strogatz network model



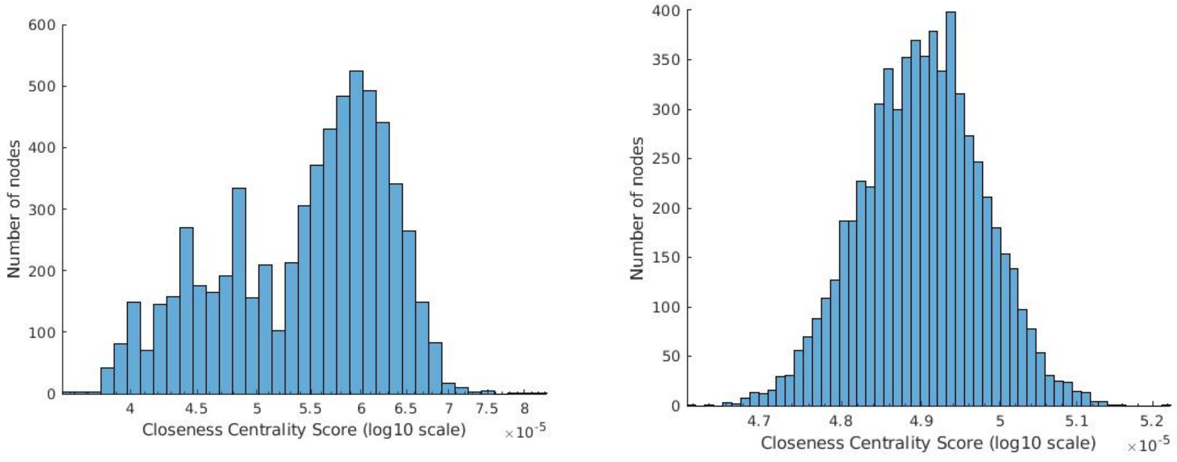
a)

b)

**Figure 3.9:** c)Betweenness centrality of a Scale-free network d) Betweenness centrality of a Random graph

The histogram of the closenness centrality has a completely different behaviour from the network models, with a local minima in the middle of the range of values. In that point few nodes have an average inverse distance to all other nodes. But looking at the overall histogram, one can observe that in general, most nodes are close to all other nodes Fig. 3.10 .

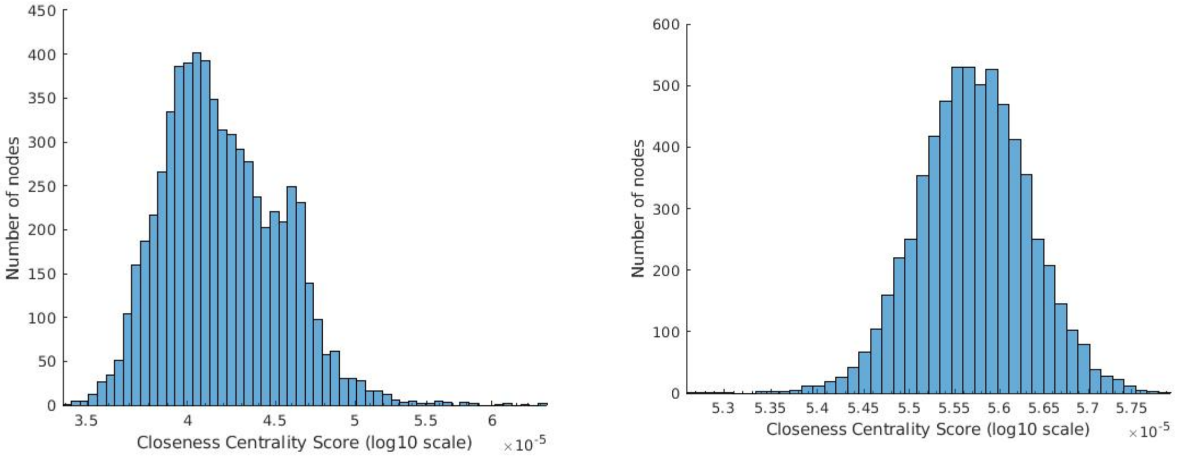
The PageRank centrality score measures the importance of each node within the graph, based on the number incoming relationships and the importance of



a)

b)

**Figure 3.10:** a) Closeness centrality of the Drug-Virus-Host interactome b) Closeness centrality of Watts-Strogatz network model

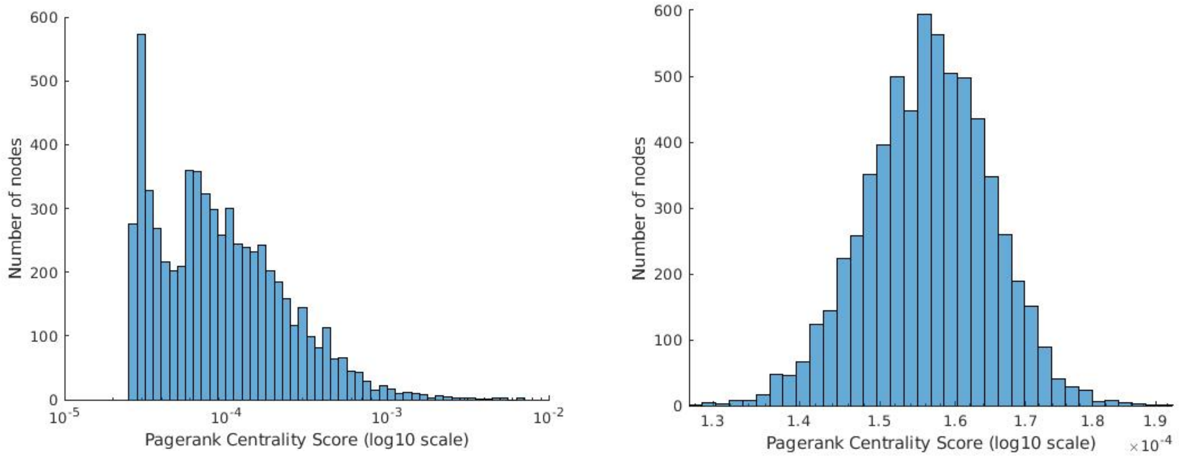


a)

b)

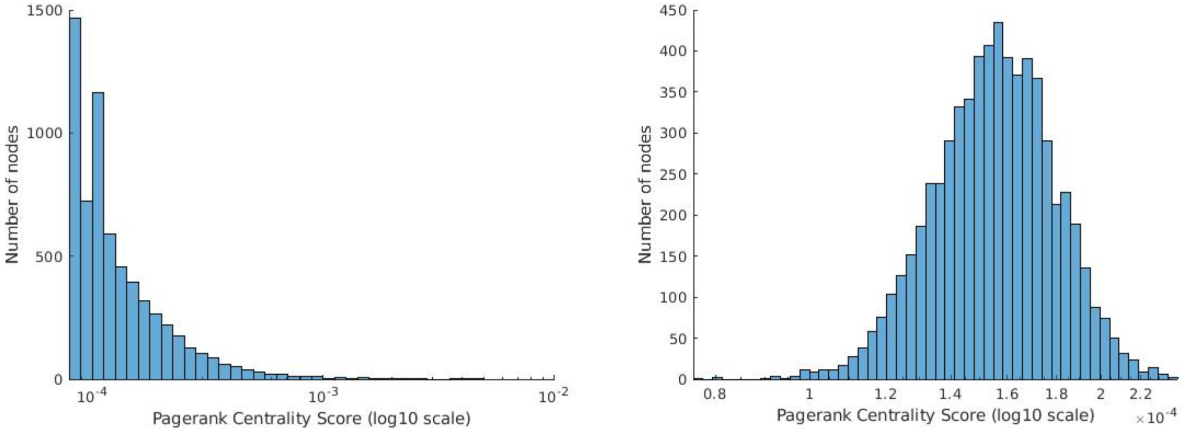
**Figure 3.11:** c) Closeness centrality of a Scale-free network d) Closeness centrality of a Random graph

the corresponding source nodes. PageRank of the interactome follows approximately a power law in the middle range of its values, in a similar way as the a scale-free growing network model of Barabasi-Albert Fig. 3.13. This fact can be interpreted as follows: few nodes has important connections.



a) b)

**Figure 3.12:** a) Pagerank centrality of the Drug-Virus-Host interactome b) Pagerank centrality of Watts-Strogatz network model



a) b)

**Figure 3.13:** c) Pagerank centrality of a Scale-free network d) Pagerank centrality of a Random graph

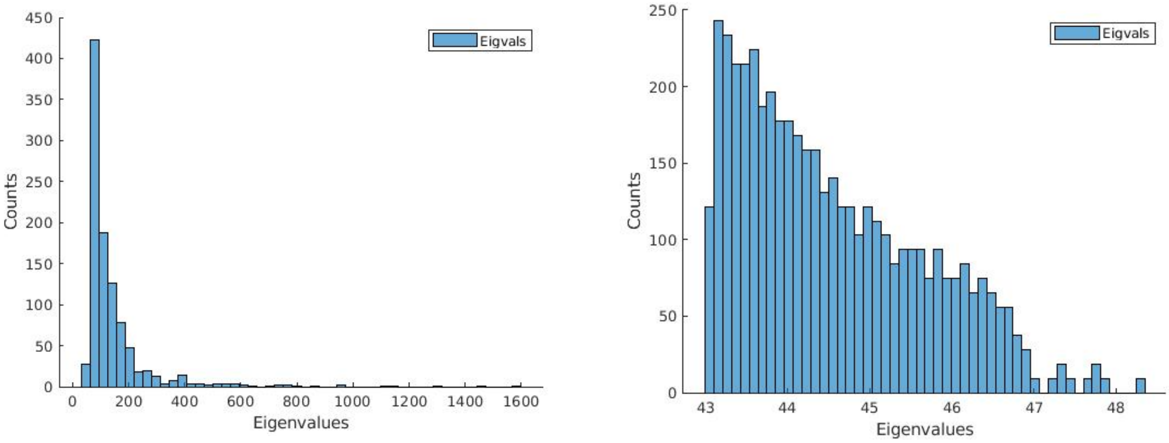
### 3.3.2 Spectral properties

After the analysis of the most common centrality measures, It's important to have a look at the properties of the Drug-Virus-Host interactome from our knowledge of its eigenvalues. The set of eigenvalues of a graph G is known as the spectrum of G and denoted by Sp(G). The spectra of the network are known to provide rich information of the topological structure and diffusion of signals. The histogram of the eigenvalues of the Laplacian displays the eigenvalues distribution. in Matlab, function eigs() return a subset of specified K eigenvalues and eigenvectors. If 'smallesttab' is specified, The first K eigenvalues and eigenvectors, sorted by ascending order, are returned. We visualize the histogram of the eigenvalues distributions obtained from the Laplacian Matrix, comparing the results with the different network models like we did for the centrality measures.

Computing the laplacian matrix in Matlab is straighforward by using

```
L = laplacian(G)
```

It's interesting to notice that the eigenvalues distribution of the interactome follows the same law as the Scale-free network. Fig. 3.15



a)

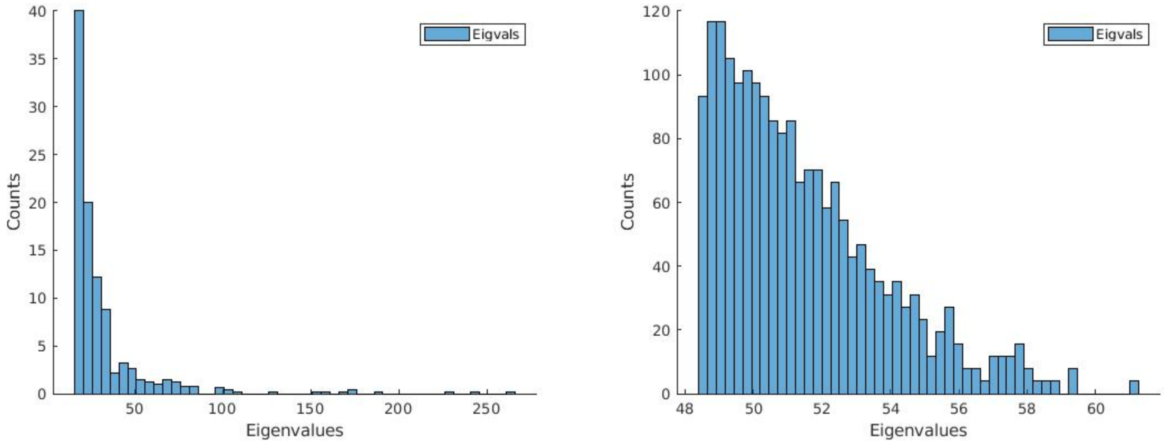
b)

**Figure 3.14:** a) Eigenvalues distribution of the Drug-Virus-Host interactome b) Eigenvalues distibution centrality of Watts-Strogatz network model

### Spectral clustering

Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories).

Spectral graph partitioning is a method of partitioning a graph into two subgraphs in such a way that the subgraphs have a balanced number of vertices (possibly equal numbers) while also minimizing the number of edges between



a) b)

**Figure 3.15:** c) Eigenvalues distibution of a Scale-free network d) Eigenvalues distibution of a Random graph

the two subgraphs.

Fiedler's theory of spectral graph partitioning uses the second eigenvector, which corresponds to the second-smallest eigenvalue, to define a semi-optimal cut to partition the vertices. This eigenvalue is called Fiedler value and the corresponding vector is called the Fiedler vector.

The second-smallest eigenvalue of $L(G)$, $\lambda 2(L(G))$, is often called the algebraic connectivity of the graph G. In general, the smaller the algebraic connectivity of a graph, the more likely it is that the Fiedler method will find a "good" partition.

According to Fiedler, the unnormalized graphs's Laplacian has the following spectral properties:

- All eigenvalues are non-negative and monotonic non-decrescent.
- If the graph is divided into K components, the multiplicity of the eigenvalue 0 of L is equal to K.
- Eigenvector components act like coordinates to represent nodes in space.
- The Fiedler vector has both positive and negative components, their sum

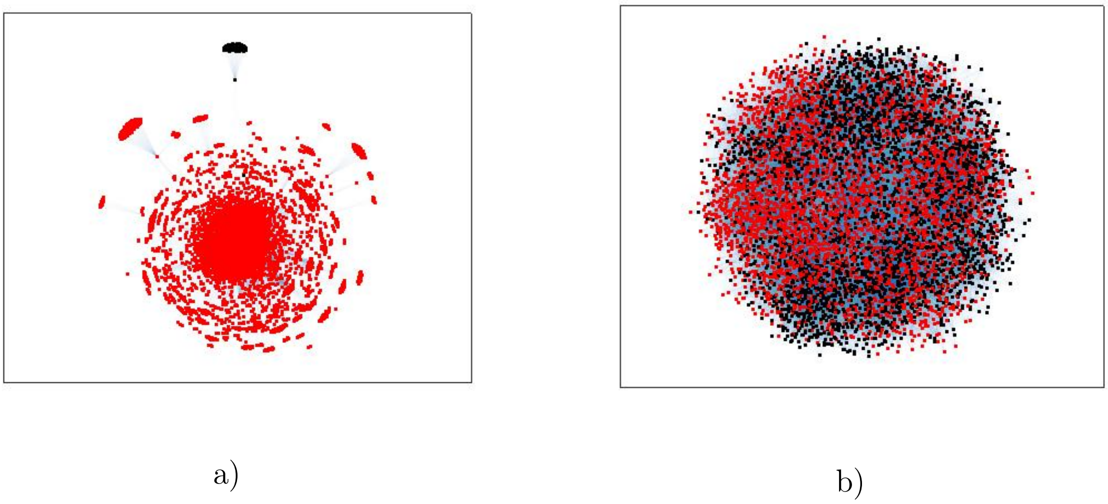
must be zero.

To make a spectral clustering using Fiedler's theory, the following procedure written in Algorithm 1 has been followed .

**Algorithm 1:** Algorithm for Graph Partitioning Using Fiedler Vector.

```
Result: Two labeled graphs, G1 and G2
Compute the Laplacian matrix;
Find the eigenvector corresponding to second smallest eigenvalue of L.;
Assign to every vertex of the original graph with label $i$ its
  corresponding entry of the Fiedler Vector.;
To form the partition, simply create two graphs G1 and G2;
for each vertex $i$ of G, if $x_i < 0$, put vertex $i$ in G1; otherwise, put it in
  G2;
```

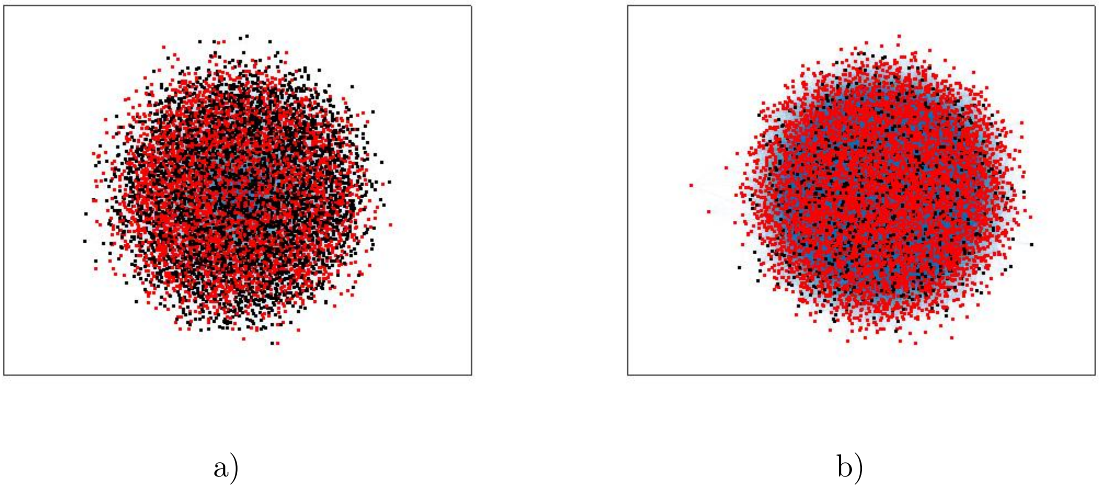
The following figures illustrate the result of the experiment.


a)

b)

**Figure 3.16:** a) Spectral clustering of the Drug-Virus-Host interactome b) Spectral clustering of Watts-Strogatz network model

Clearly, the cut is unbalanced in the case of the interactome Fig. 3.16 a). This is not a suprise: we have only one null eigenvalue. The number of null eigenvalues is a good indicator of the number of connected components in a similarity graph and, therefore, is a good estimate of the number of clusters in the data. These plots confirms the theory. The cluster with black nodes is composed by only one human protein, DPP4, one drug, Atorvastatin, both



a) b)

**Figure 3.17:** c) Spectral clustering of a Scale-free network d) Spectral clustering of a Random graph

with a degree greater tha 1. The rest of the nodes in the black colored cluster are drugs connected to DPP4. This result is interesting because the statin class of drugs as showed benefits in the treatment of COVID-19. I will talk about it in more details in the next paragraphs [25].

3.16

### Other metrics and distributions

In Fig. 3.18 there is a visualization of cumulative distance distribution data of the interactome. The x axis represents distance values while the y axis represents percent of node pairings that are connected by that on a shorter distance. In Fig. 3.19 there is a visualization of cumulative local clustering coefficient distribution data of the interactome. The x axis represents local clustering coefficient values while the y axis represents percent of nodes that have that or a lesser local clustering coefficient (hence cumulative percentage).
Finally, in Fig. 3.20 there is a visualization of assortativity data of the interactome. The x axis represents node degree values while the y axis represents the average degree of neighboring nodes. Both axes use a logarithmic scale. A scatter plot shows average degree of neighbors on a node-by-node basis, whereas

a line shows the average degree of neighbors for all nodes of a specific degree. In essence, the points of the line merely depict a weighted average (vertically) of the scatter plot.

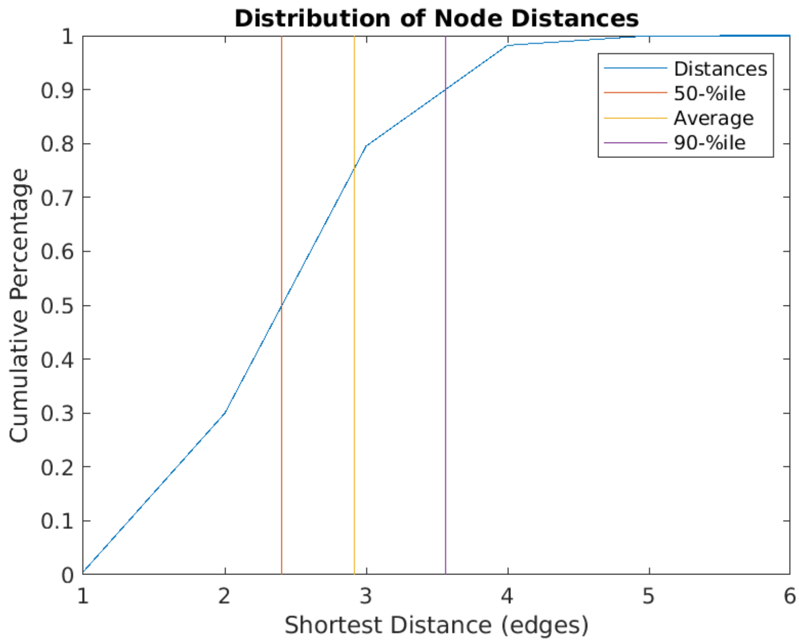


**Figure 3.18:** Distance distribution of node distance in the interactome

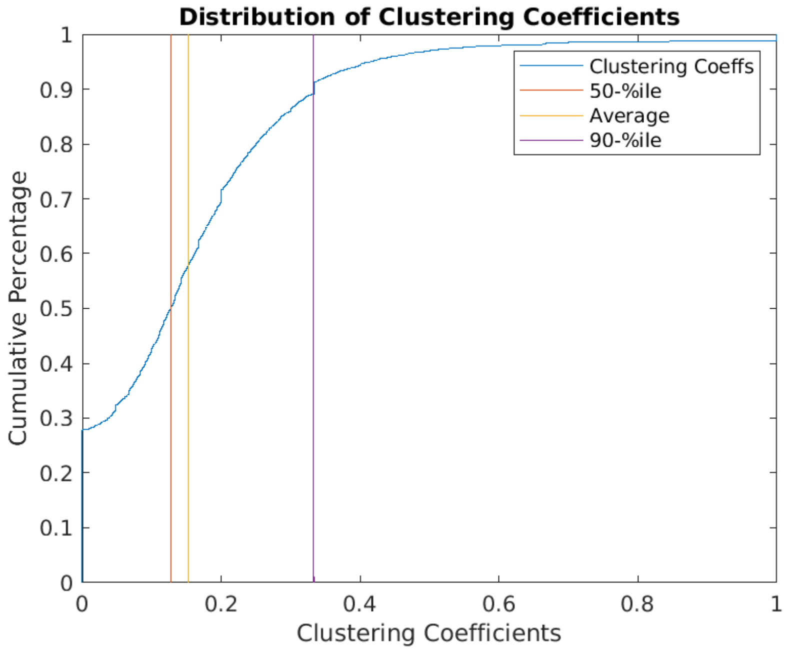


**Figure 3.19:** Distribution of clustering coefficients in the interactome

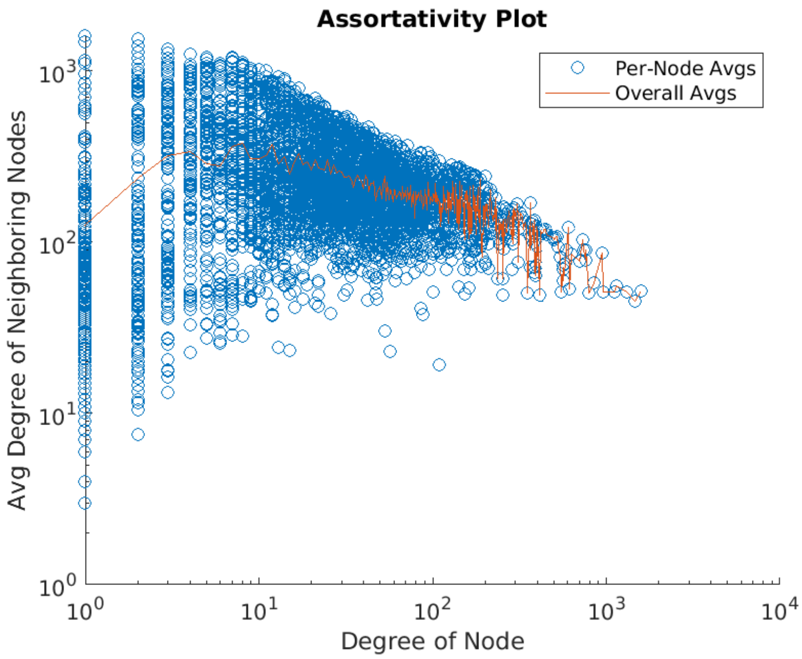


**Figure 3.20:** Assortativity plot

## 3.4 Drug Prioritization Strategy

To propose drugs to be repurposed for the treatment of COVID-19 we need to infer and prioritize high ranking genes that may be potential pharmacological targets.

The basis for network-based drug repurposing rests on the notions that:

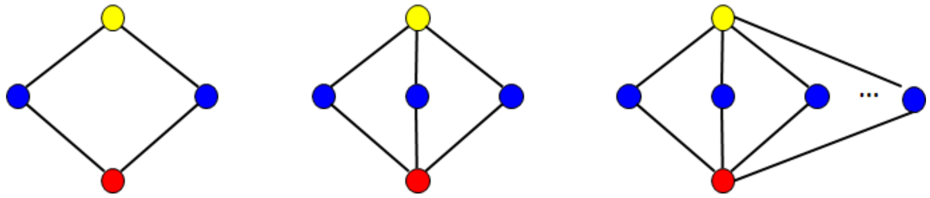
- The proteins that associate with and functionally govern a disease phenotype are localized in the corresponding disease module or sub-network within the comprehensive PPI network.

- Proteins that serve as drug targets for a specific disease may also be suitable drug targets for another disease owing to common PPIs and functional pathways elucidated by the PPI [26].

Most antiviral drugs target the virus proteins or their direct host interactor proteins to inhibit different stages of the viral life cycle.

The intuition behind this work is on selecting drug candidates targeting human proteins which interacts with the same viral proteins. We aim to identify repurposable drug candidates with highest out degree, which significantly perturb the disease module. At the same time, we have under control functional modules that are responsible for the virus replication and the disease progression following infection. In this way we are combining both the benefits of targeting the human host proteins directly responsible for the disease while keeping the network-based considerations in mind. This is especially useful when we have poor knowledge of the disease module.

Based on the previous consideration, network motifs were sought that followed a diamond pattern. Formally, we are searching for putative drugs which forms the diamond-shape motifs in the tripartite graph of the Virus-Host-Host-Drug network like that on the figure.



The algorithm, written in Matlab, is straighforward: for every node that represent a drug (yellow) and for every Sars-Cov-2 node (red) the drug-targets is the intersection of the sets of the two neighborhoods.

**Algorithm 2:** Search of diamond-shape network motif

```
Result: List of diamond shape network motifs
initialization;
while While condition do
    select a node drug;
    return a list of drug's neighbors;
    while While condition do
        select a sars-cov-2 node;
        return a list of sars neighbors;
        find the insercection between the two neighbors;
        write to file a list of: drug, intersection neighbors, sars-cov-2
          gene;
    end
end
```

## 3.5 Functional enrichment

A powerful analytical method for interpreting gene expression data is called Gene Set Enrichment Analysis (GSEA) [27]. The method focuses on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. Gene Set Enrichment Analysis is able to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes.

The Gene Ontology (GO) is the concept of associating a collection of genes with a functional biological term in a systematic way. The creation of GO enabled the analysis of gene lists in the context of prior knowledge.

Given a ranked list of genes L, taken from a high-throughput experiment, the challenge is to extract meaning from this list. Given an a priori defined set of genes S (known genes sharing the same GO category or involved in the same metabolic pathway), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. there are three steps involved in the analytical process:

- Calculation of an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated as a weighted Kolmogorov–Smirnov-like statistic.
- Estimate the statistical significance of the ES. This calculation is done by a permutation test in order to produce a null distribution for the ES. The P value is determined by comparison to the null distribution (How unlikely such a result is to occur by chance?).

For the task of functional analysis Enrichr was used. Enrichr is an easy to use intuitive enrichment analysis web-based tool providing various types of visualization summaries of collective functions of gene lists. Enrichr is open source and freely available online at: `http://amp.pharm.mssm.edu/Enrichr`. Enrichr currently contains a large collection of diverse gene set libraries available for analysis and download.
Enrichr uses a scoring method similar to the Hypergeomtric/Fisher's exact test. To do such a test, we need to have a list of interesting (e.g. differentially expressed) genes and a list of the background (also known as universe). However, Enrichr has its own background lists, so we do not need to specify them explicitly.

### 3.5.1 Centrality-based Enrichment

For every centrality measure, a list of top human genes is retrieved and then passed to Enrichr to perform the enrichment analysis. The intuition behind this

methodology is that different centrality measurements may relate to different biological functions.
For every centrality measures it was retrieved:

- **GO Biological Process** is the ontology that covers operations or sets of molecular events with a defined beginning and end. Examples of broad biological process terms are DNA repair or signal transduction.
- **GO Molecular Function** covers molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”.
- **GO Cellular Component** refers to the locations relative to cellular structures in which a gene product performs a function.
- **COVID-19** related gene sets.
- **KEGG 2019 Human**. The Kyoto Encyclopedia of Genes and Genomes (KEGG), is a database for understanding high-level functions and utilities of the biological system. It provides comprehensible manually-drawn pathways representing biological processes or disease-specific pathways.

### Top proteins by Degree

To perform this enrichment analysis, the top 100 human gene symbols ranked by degree was given as input to Enrichr. Looking at 3.26 and 3.24, a series of facts can be deduced:

- there is an over representation of genes involved in liver and lung carcinoma as well as endometrial and prostate cancer. In fact, there is a correlation with the clinical observations in which cancer patients generally have a higher risk of infection and worse outcome.

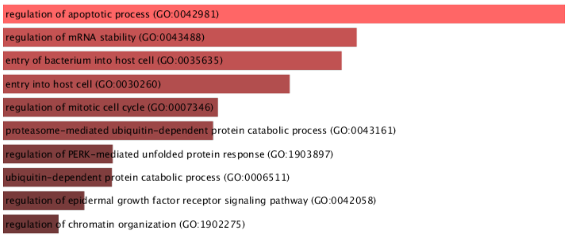


**Figure 3.21:** GO Biological Process

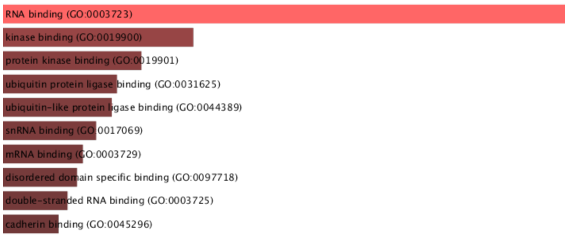


**Figure 3.22:** GO Molecular Function

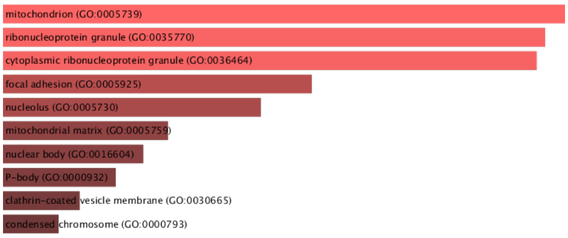


**Figure 3.23:** GO Cellular Component

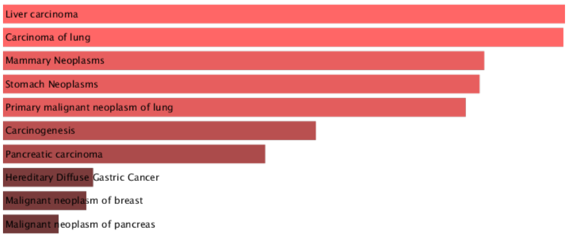


**Figure 3.24:** DisGeNET diseases

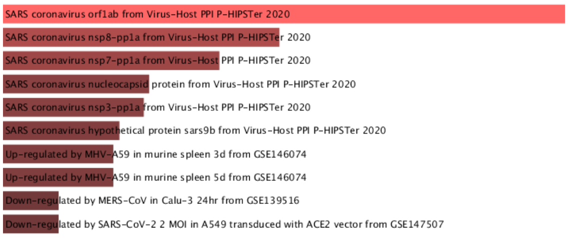


**Figure 3.25:** COVID-19 related gene sets

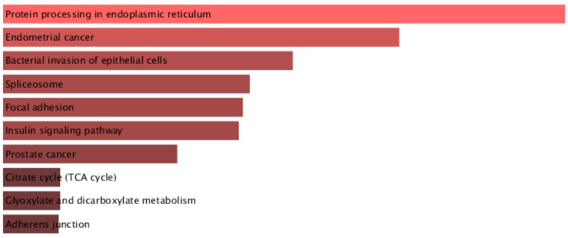


**Figure 3.26:** KEGG 2019 Human Pathway

### Top proteins by betweenness

Top ranked proteins by betweenness centrality act as a bridge in the interactome.

- Most of these proteins are involved in the interaction with the SARS protein orf1ab.
- This proteins are mostly involved in lung carcinoma, has a pathway in cancer.
- positive regulation of gene expression indicates any process that increases the frequency, rate or extent of gene expression. This can be true in case of viral activity.

- The antigen processing and presentation is at the core of immunological process.

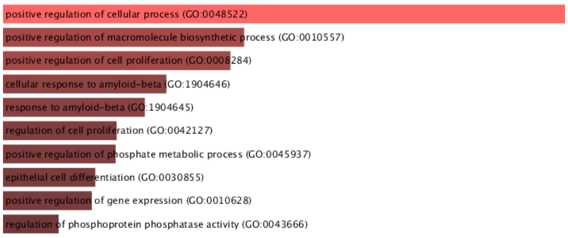


**Figure 3.27:** GO Biological Process

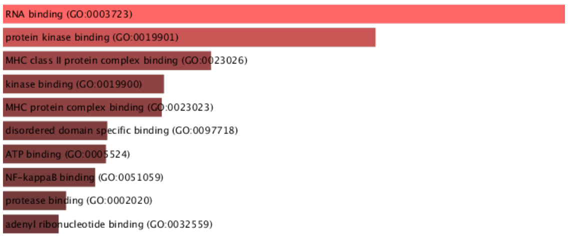


**Figure 3.28:** GO Molecular Function

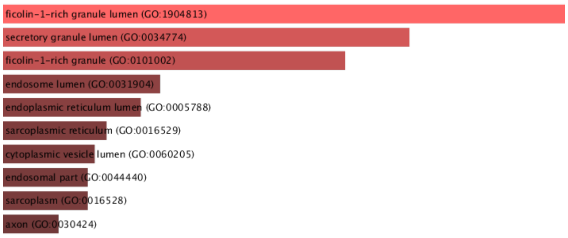


**Figure 3.29:** GO Cellular Component

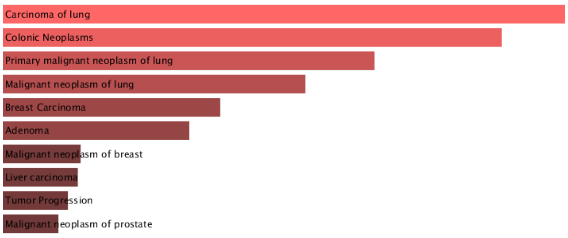


**Figure 3.30:** DisGeNET diseases

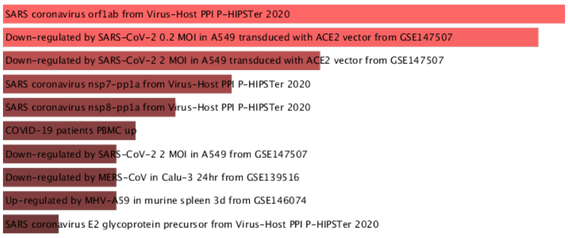


**Figure 3.31:** COVID-19 related gene sets

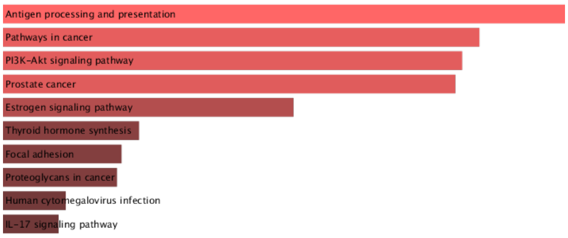


**Figure 3.32:** KEGG 2019 Human Pathway

### 3.5.2 Top proteins by closeness

- Most of these proteins are involved in the interaction with the SARS protein nsp8.

- The first pathway score is different from that of top degree and betweenness, although the pathway in cancer is still represented.

- The most over-represented disease associated to this gene set is melanoma.

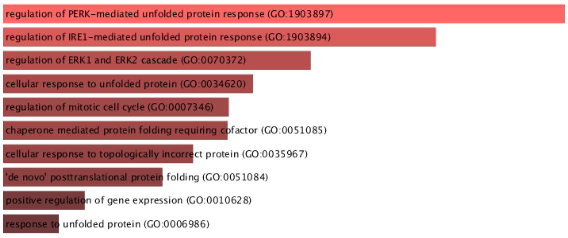


**Figure 3.33:** GO Biological Process

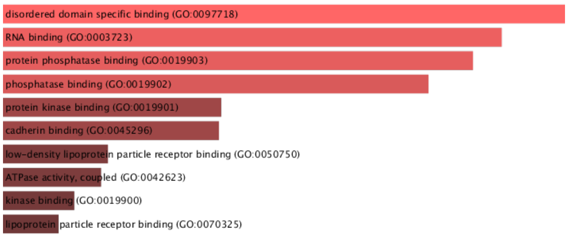


**Figure 3.34:** GO Molecular Function

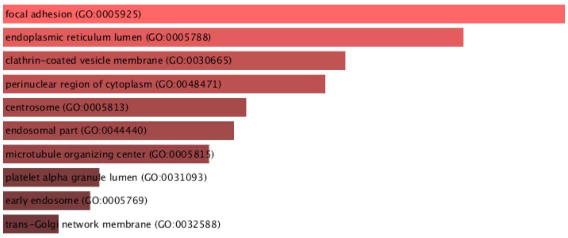


**Figure 3.35:** GO Cellular Component

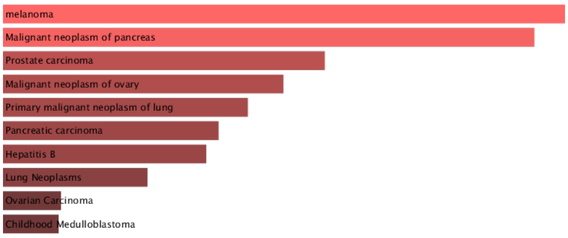


**Figure 3.36:** DisGeNET diseases

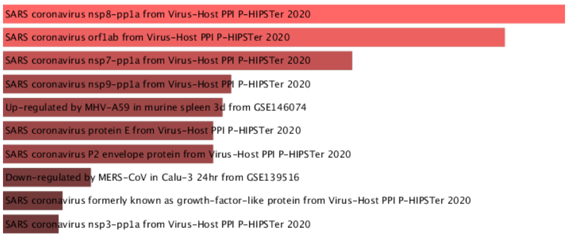


**Figure 3.37:** COVID-19 related gene sets

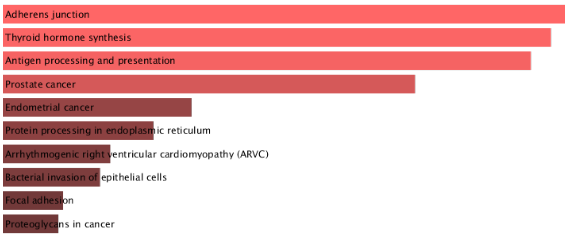


**Figure 3.38:** KEGG 2019 Human Pathway

### Top proteins by pagerank

- Most of these proteins are involved in the interaction with the SARS protein nsp8.
- The most over-represented disease in DisGeNET is melanoma. The perturbation of the top pagerank genes are also involved in other cancer-related diseases.
- The biological process is that of bacteria that enters a host cell.
- The molecular function is RNA and kinase binding.
- The pathway is that of adherens junction.

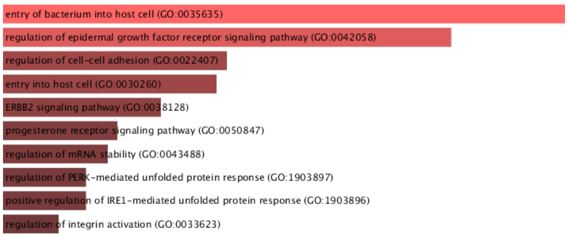


**Figure 3.39:** GO Biological Process

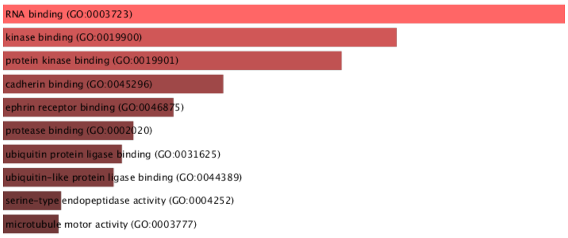


**Figure 3.40:** GO Molecular Function

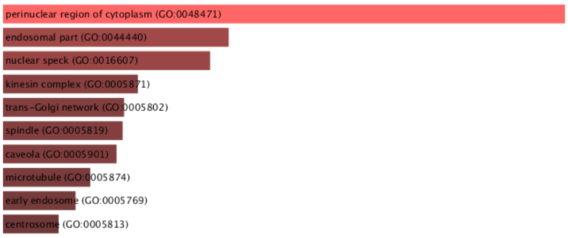


**Figure 3.41:** GO Cellular Component

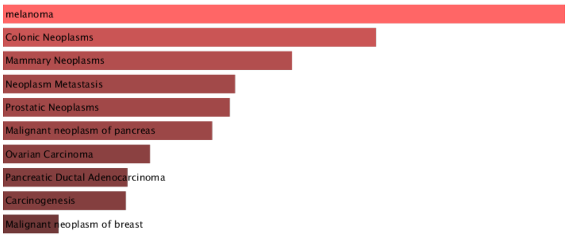


**Figure 3.42:** DisGeNET diseases

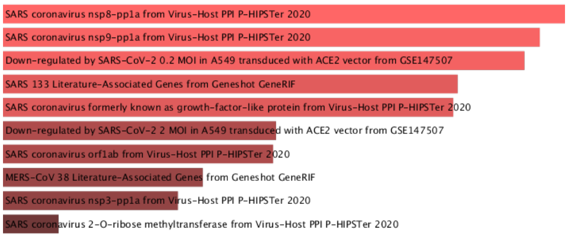


**Figure 3.43:** COVID-19 related gene sets

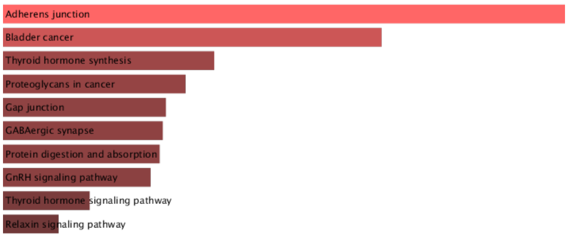


**Figure 3.44:** KEGG 2019 Human Pathway

### Top proteins by eigenvector

- The genes from this list seems to be more related to the previous MERS-CoV.
- We have an increased of serum pyruvate.
- The biological process is that of sulfur amino-acid biosynthesis
- The molecular function is that of the ATP binding.
- The pathway is that of cardiac muscle contraction.

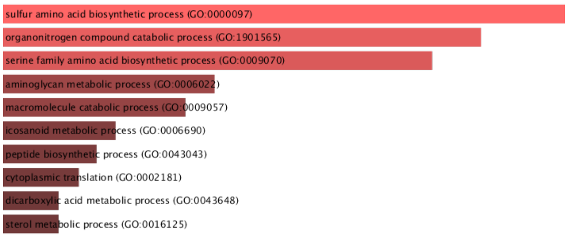


**Figure 3.45:** GO Biological Process

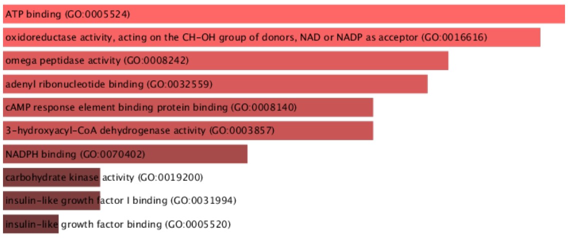


**Figure 3.46:** GO Molecular Function

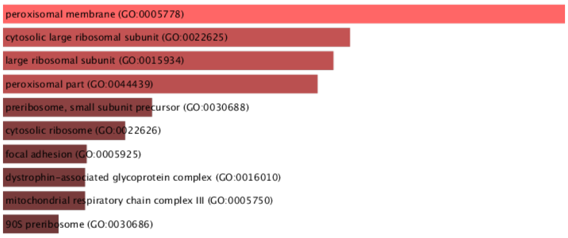


**Figure 3.47:** GO Cellular Component

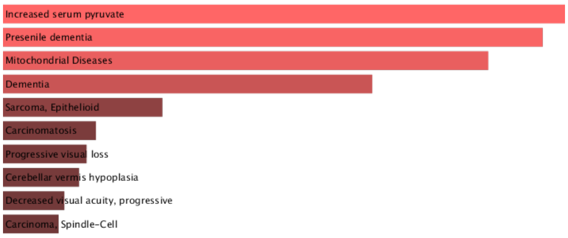


**Figure 3.48:** DisGeNET diseases

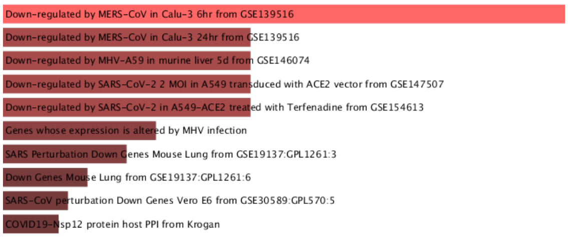


**Figure 3.49:** COVID-19 related gene sets

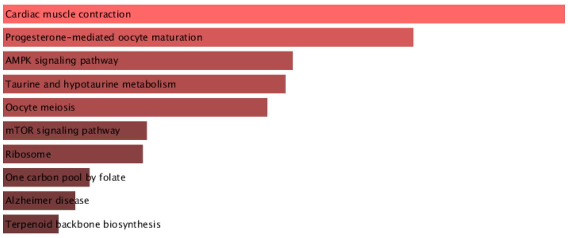


**Figure 3.50:** KEGG 2019 Human Pathway

# Chapter 4

# Results

The Tab.4.1 shows the results of repurposed drugs discovered from the diamond-shape motifs, ordered in descending order by the number of binding targets from the Virus-host interactome.

Several examples exist of repurposed drugs being or having been tested in clinical trials for COVID-19, including antiviral drugs and host-targeting therapies.

*Irinotecan* is an antineoplastic enzyme inhibitor primarily used in the treatment of colorectal cancer. Irinotecan was approved for the treatment of advanced pancreatic cancer in October, 2015. This combination could potentially be an exceptionally effective treatment to protect critically ill patients from death caused by COVID-19-specific cytokine storms triggered by sepsis, ARDS, and other fatal comorbidities [28].

The algorithm discovered several drugs from the *statin* class of medications. A molecular docking study was made to investigate if statins could be efficient SARS-CoV-2 Mpro inhibitors [29]

.

Surprisingly, nutraceuticals such as copper, fatty acids and retinoils, can also be used as an aid in prevention of the Covid-19. In fact, functional foods and food supplements within popular diets contain immune-boosting nutraceuticals, polyphenols, flavonoids, alkaloids, used to optimize the immune system capacity to prevent and control pathogenic viral infections, while physical activity augments such protective benefits [30].

The algorithm discovered Nicotinamide adenine dinucleotide (NAD) as a potential treatment, which is a cofactor central to metabolism. Administration of anti-ageing immunomodulation factors like NAD+ can minimize the severity of the COVID-19 symptoms in elderly people through its potent immunomodulation and longevity effects [31].

From the class of Antiviral drugs it appears Lamivudine which is used to treat HIV-1 and hepatitis B (HBV).

From the common medications worth of mention is Ibuprofen and Paracetamol. In vitro evidence suggests that Ibuprofen may have a role in reducing excess inflammation or cytokine release in COVID-19 patients [32]. Other studies suggests that, in case of infection, symptomatic treatment with NSAIDs - like Ibuprofen - for non-severe symptoms (fever, pain, or myalgia) is not to be recommended, given a range of clinical and scientific arguments supporting an increased risk of severe bacterial complication. Besides, paracetamol at recommended doses, is a safer alternative [33].

## 4.1 Related Work

Until now, computational drug-repositioning approaches applied on COVID-19 can be broadly categorized as network-based models, structure-based ap-

| **Drug** | **Function** |
|---|---|
| Irinotecan | Anti-tumoral |
| Paracetamol | Common Analgesic |
| Tretinoin | Tretinoin is a naturally occurring derivative of vitamin A (retinol) |
| NADH | NADH is the reduced form of NAD+, a coenzyme. |
| Copper | Copper is commonly used in contraceptive intrauterine devices (IUD) |
| Alitretinoin | Tretinoin, or retinoic acid and derived from maternal vitamin A |
| Zoledronic acid | It is used to treat and prevent multiple forms of osteoporosis. Anti-tumoral. |
| Simvastatin | It belongs to the statin class of medications. |
| Rosuvastatin | It's a lipid-lowering drug that belongs to the statin class of medications. |
| Risedronic acid | It is used for the treatment of some forms of osteoperosis and Paget's disease. |
| Pravastatin | It belongs to the statin class of medications. |
| Pamidronic acid | Is a nitrogen containing bisphosphonate that inhibits osteoclast mediated bone loss. |
| Palmitic Acid | A common saturated fatty acid. Natural integrator. |
| Morphine | It is a potent analgesic. |
| Lovastatin | It belongs to the statin class of medications. |
| Lamivudine | Antiviral. It is used to treat (HIV-1) and hepatitis B (HBV). |
| Isoprenaline | It is used mainly as bronchodilator and heart stimulant. |
| Ibuprofen | It is a non-steroidal anti-inflammatory drug (NSAID). |
| Ibandronate | It is used to prevent and treat postmenopausal osteoporosis. |
| Fluvastatin | Statin-class medication. |
| Etoposide | Anti-tumoral. |
| Doxorubicin | Is cytotoxic anthracycline antibiotic. |

**Table 4.1:** List of drugs resulting from the motif-based prioritization method

proaches or machine/deep learning approaches.

As typically only a small subset of the top candidates is validated experimentally, the true predictive power of the existing repurposing algorithms remains unknown.

In this study [34], network proximity between drug targets and virus associated proteins was calculated to screen for candidate repurposable drugs under the human protein interactome model and then the "Complementary Exposure" pattern was applied to find combination of therapies: the targets of the drugs both hit the Virus-host subnetwork, but target separate neighborhoods in the human interactome network.

Another interesting network-based framework combines three different approaches, resulting in 12 different pipelines for drug ranking [35]. They implemented three competing network repurposing methodologies:

- the AI based, in which they designed a graph neural network for COVID-19 treatment recommendations based on a graph neural network (GNN) architecture.
- The diffusion algorithm ranks drugs based on capturing network similarity of a drug's protein targets to the SARS-CoV-2 host protein targets.
- The proximity algorithm ranks drugs based on the distance between the host protein targets of SARS-CoV2 and the closest protein targets of drugs.

Methods that are based on graph embedding have been gaining recent attention for link prediction in graphs that represent nodes and edges as low-dimensional feature vectors. Using the feature vectors of drugs and diseases, it is easy to measure their similarities and therefore identify effective drugs for a given disease.

In particular, medical knowledge graphs have been used for drug repurposing [36]. An example of existing toolbox is the Drug Repurposing Knowledge Graph (DRKG) which is a comprehensive biological knowledge graph relating genes, compounds, diseases, biological processes, side effects and symptoms. DRKG includes information from six existing databases including DrugBank, Hetionet, GNBR, String, IntAct and DGIdb, and data collected from recent publications particularly related to Covid-19 [37].
The usage of multiple drugs offers an increased therapeutic efficacy and reduced toxicity, play an important role in treating infectious diseases, including COVID-19.
However, the ability to identify and validate effective combinations is limited due to the huge space of drug-drug combinations. Nowadays, drug combinations are based on theoretical analysis using the interactome and have not been tested in preclinical or clinical studies.

## 4.2 Future Work

### 4.2.1 Discovery of network motifs & functional analysis

To evaluate functional pathways of proteins involved in the diamond shape network motifs found in the virus-host interactome, it is essential to perform gene enrichment analysis using Kyoto Encyclopedia of Genes and Genomes (KEGG) human pathways and Gene Ontology databases [38], as it can help to elucidate the mechanism of action of the drugs.

In the future, I want to find more network motifs within these diamond-shaped submodules. I may decide to prioritize a drug within the diamond-shape pattern, within which human proteins form a triangle pattern. This motif consists of three nodes and is frequently encountered in regulatory networks. Authors of this study [39], proves that triangle shaped network motifs are indeed enriched with drug targets and can serve as potential targets.

### 4.2.2 Drug combinations for COVID-19

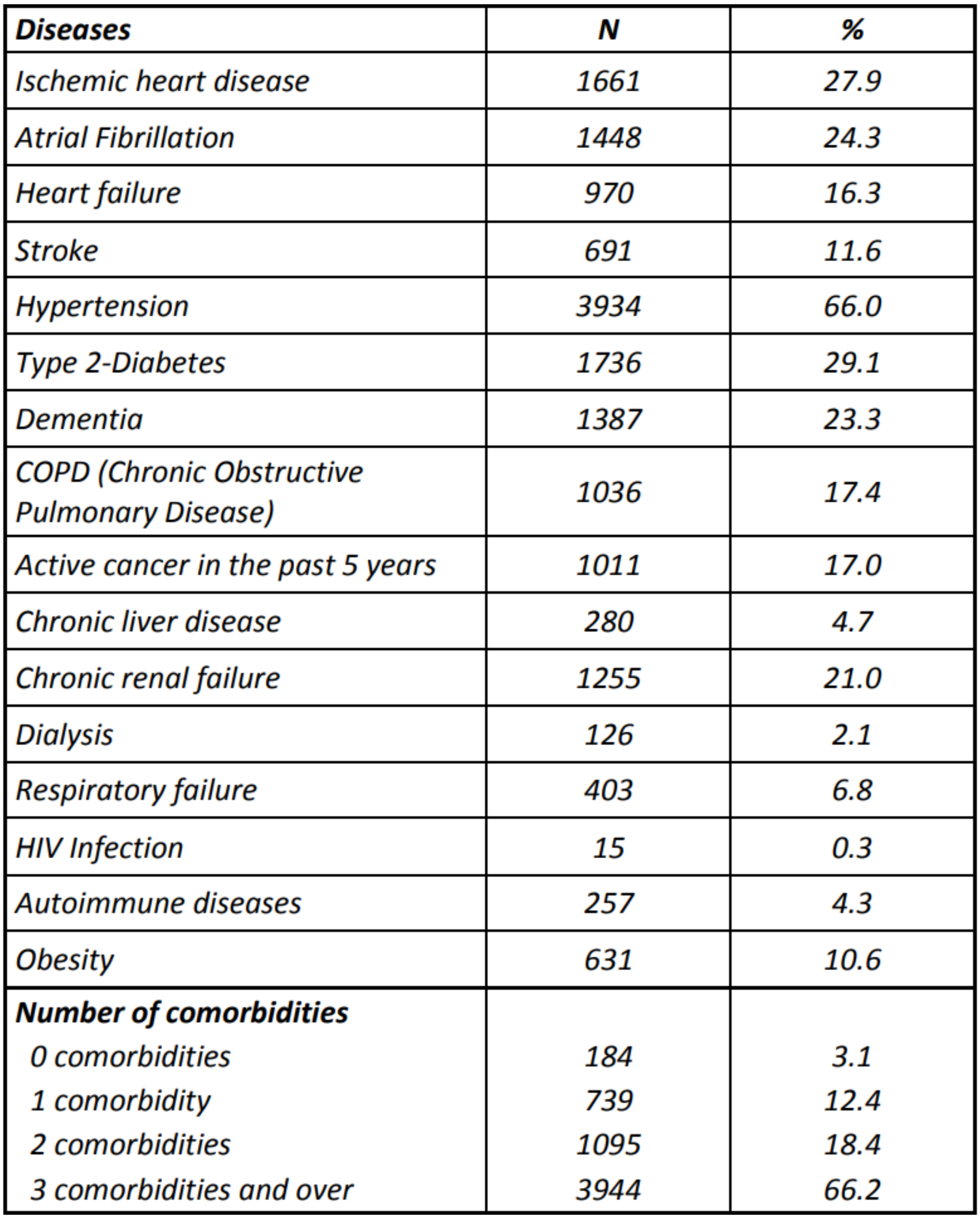
An interesting follow-up of the research is considering drug combinations. The implementation could be similar to that proposed in [40], where a network-based methodology was proposed to identify clinically efficacious drug combinations for specific diseases. They show the existence of six distinct classes of drug–drug–disease combinations finding that only one of the six classes correlates with therapeutic effects: if the targets of the drugs both hit disease module, but target separate neighborhoods.
It could be interesting to integrate their results in this work, by finding classes of drug-drug combinations on network motifs. Intuitively, the class with therapeutic effect could be that of separate motifs and different neighborhood but in the same disease module.

### 4.2.3 The need for a personalized drug repositioning

SARS-CoV-2 infection has shown large inter-individual variabilities, ranging from asymptomatic to severe and lethal disease, although clinical data show some patterns that can be exploited.
From the report "Characteristics of COVID-19 patients dying in Italy" by the italian association for public health Istituto Superiore di Sanità:

> *“Before hospitalization, 21% of SARS-CoV-2 positive deceased patients followed ACE-inhibitor therapy and 14% angiotensin receptor blockers-ARBs therapy. This information can be underestimated*

*because data on drug treatment before admission were not always described in the chart.”*



| Diseases | N | % |
|---|---|---|
| Ischemic heart disease | 1661 | 27.9 |
| Atrial Fibrillation | 1448 | 24.3 |
| Heart failure | 970 | 16.3 |
| Stroke | 691 | 11.6 |
| Hypertension | 3934 | 66.0 |
| Type 2-Diabetes | 1736 | 29.1 |
| Dementia | 1387 | 23.3 |
| COPD (Chronic Obstructive Pulmonary Disease) | 1036 | 17.4 |
| Active cancer in the past 5 years | 1011 | 17.0 |
| Chronic liver disease | 280 | 4.7 |
| Chronic renal failure | 1255 | 21.0 |
| Dialysis | 126 | 2.1 |
| Respiratory failure | 403 | 6.8 |
| HIV Infection | 15 | 0.3 |
| Autoimmune diseases | 257 | 4.3 |
| Obesity | 631 | 10.6 |
| **Number of comorbidities** | | |
| 0 comorbidities | 184 | 3.1 |
| 1 comorbidity | 739 | 12.4 |
| 2 comorbidities | 1095 | 18.4 |
| 3 comorbidities and over | 3944 | 66.2 |

**Figure 4.1:** Most common comorbidities observed in SARS-CoV-2 positive deceased patients

Antibiotics were used by 85.8% of patients during hospital stay, while less used were corticosteroids (51.3%) and antivirals (48.7%). Concomitant use of these 3 treatments was observed in 23.9% of cases. Out of SARS-CoV-2 positive deceased patients, 4.1% were treated with Tocilizumab during hospitalization.

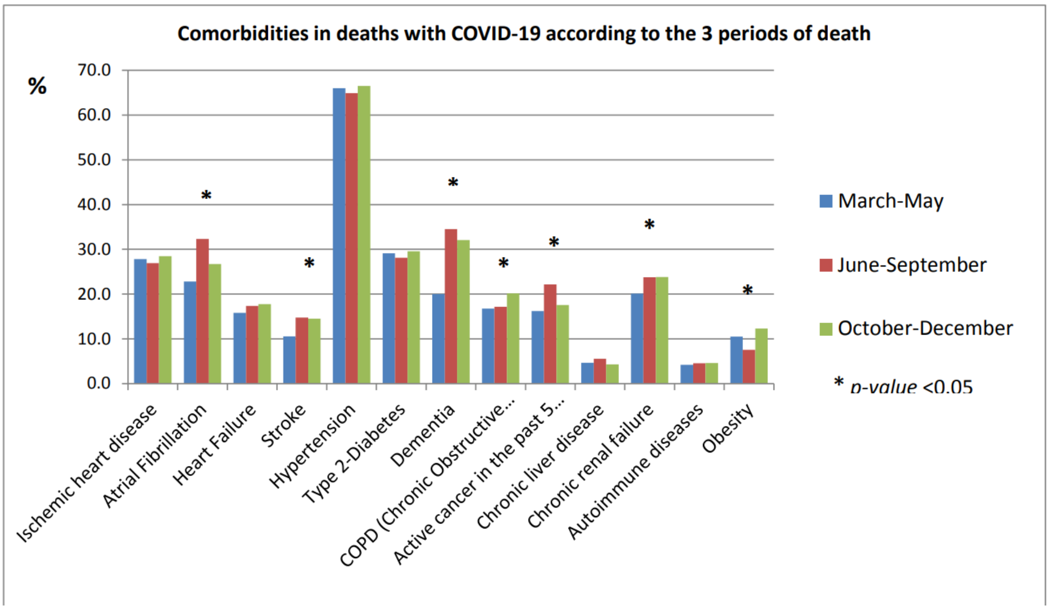


**Figure 4.2:** Mean age, prevalence of women, number of pre-existing diseases, complications and treatments in deaths with COVID-19 in the 3 periods March-May, June- September, and October-16th December 2020

We can deduce these facts:

- most of the COVID-19 related death were due to specific comorbidies.
- Even if there is no data about the previous therapies from the diseases one can infer that patients were already following specific therapy

One strategy could be that of building the interactome composed by the disease modules taken from the list of comorbidities with COVID-19, and then looking at the drug–drug–disease combinations. It can be followed a network-based strategy or a more sophisticated graph neural network.

We expect future computational methods for drug repurposing will also take into consideration personal genetic and clinical data. COVID-19 disease treatment would be considerably improved if therapies were guided by individual's genomic profiles.

This study confirms the clinical observations [41] that ACE2 or TMPRSS2 DNA polymorphisms were likely associated with genetic susceptibility of COVID-19, which calls for a human genetics initiative for fighting the COVID-19 pandemic.

AI techniques could leverage massive genetic and genomic data for drug repurposing and personalised treatment for individuals with COVID-19.

# Chapter 5

# Conclusions

This work presented a computational, network-based framework for drug repurposing. The focus was the search for a drug treatment for COVID-19, using a technique based on the enumeration, folowed by a prioritization, of drugs based on diamond-shape network motifs.
Before the repurposing steps, I spent a significant amount of effort on centrality measures of the interactome obtained from the gene interaction between the SARS-CoV-2 organism, the human organism and drugs. The goal of this analysis was to show the overall centrality measure distribution and extract top-ranked proteins to later perform a functional analysis using *Enrichr*. The spectral analysis shows that the spectra of the interactome follows a scale-free distribution. Then, a spectral clustering was performed by using the Fiedler vector. The results shows that the graph is connected, with just a small cluster isolated from the rest of the network.

Most of the drugs resulting from the prioritization have been widely discussed in the medical literature as a possible treatment for COVID-19. Although the results are intriguing, further investigation and work may lead to more robust results.

# Bibliography

[1] Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, Feb 2004.

[2] Albert-Laszlo Barabasi, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, Jan 2011.

[3] Jack W. Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, Mar 2012.

[4] Gina Kolata. *U.S. APPROVES SALE OF IMPOTENCE PILL; HUGE MARKET SEEN.*

[5] G. J. Morgan and F. E. Davies. Role of thalidomide in the treatment of patients with multiple myeloma. *Crit Rev Oncol Hematol*, 88 Suppl 1:14–22, Oct 2013.

[6] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong, and G. Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res*, 30(3):269–271, 03 2020.

[7] Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4(1):213, January 2008.

[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002.

[9] E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8):e1000140, Aug 2008.

[10] Andrew X. Chen, Christopher J. Zopf, Jerome Mettetal, Wen Chyi Shyu, Joseph Bolen, Arijit Chakravarty, and Santhosh Palani. Scale-free structure of cancer networks and their vulnerability to hub-directed combination therapy. *bioRxiv*, 2020.

[11] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):e59, April 2007.

[12] H.-W. Ma and A.-P. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430, July 2003.

[13] Xiao-Dong Zhang. The Laplacian eigenvalues of graphs: a survey. *arXiv e-prints*, page arXiv:1111.2897, November 2011.

[14] P Erdös and A Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[15] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

[16] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[17] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[18] Weitz Group at Georgia Tech. *GraphLab, Brighton 2020*.

[19] J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, Jun 2010.

[20] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–539, Jan 2006.

[21] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–455, Jan 2004.

[22] Emilie Chautard, Marie Fatoux-Ardore, Lionel Ballut, Nicolas Thierry-Mieg, and Sylvie Ricard-Blum. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Research*, 39(suppl$_1$) $: D235 - -D240, 092010$.

[23] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–

a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–416, Jan 2009.

[24] L. Yao and A. Rzhetsky. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res*, 18(2):206–213, Feb 2008.

[25] W. Y. T. Tan, B. E. Young, D. C. Lye, D. E. K. Chew, and R. Dalan. Statin use is associated with lower disease severity in COVID-19 infection. *Sci Rep*, 10(1):17458, 10 2020.

[26] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A. L. Barabasi. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, Feb 2015.

[27] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[28] B. Lovetrue. The AI-discovered aetiology of COVID-19 and rationale of the irinotecan+ etoposide combination therapy for critically ill COVID-19 patients. *Med Hypotheses*, 144:110180, Nov 2020.

[29] ?. Reiner, M. Hatamipour, M. Banach, M. Pirro, K. Al-Rasadi, T. Jamialahmadi, D. Radenkovic, F. Montecucco, and A. Sahebkar. Statins and the COVID-19 main protease: in silico evidence on direct interaction. *Arch Med Sci*, 16(3):490–496, 2020.

[30] A. Alkhatib. Antiviral Functional Foods and Exercise Lifestyle Prevention of Coronavirus. *Nutrients*, 12(9), Aug 2020.

[31] H. M. Omran and M. S. Almaliki. Influence of NAD+ as an ageing-related immunomodulator on COVID 19 infection: A hypothesis. *J Infect Public Health*, 13(9):1196–1201, Sep 2020.

[32] L. Smart, N. Fawkes, P. Goggin, G. Pennick, K. D. Rainsford, B. Charlesworth, and N. Shah. A narrative review of the potential pharmacological influence and safety of ibuprofen on coronavirus disease 19 (COVID-19), ACE2, and the immune system: a dichotomy of expectation and reality. *Inflammopharmacology*, 28(5):1141–1152, Oct 2020.

[33] J. Micallef, T. Soeiro, and A. P. Jonville-B?ra. Non-steroidal anti-inflammatory drugs, pharmacology, and COVID-19 infection. *Therapie*, 75(4):355–362, 2020.

[34] Yadi Zhou, Yuan Hou, Jiayu Shen, Yin Huang, William Martin, and Feixiong Cheng. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell Discovery*, 6(1):14, Mar 2020.

[35] Gysi Deisy Morselli, Do Valle Ítalo, Zitnik Marinka, and Barabási Albert-László. Network medicine framework for identifying drug repurposing opportunities for covid-19.

[36] Y. Zhu, C. Che, B. Jin, N. Zhang, C. Su, and F. Wang. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics J*, 26(4):2737–2750, Dec 2020.

[37] Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. Drkg - drug repurposing knowledge graph for covid-19. `https://github.com/gnn4dr/DRKG/`, 2020.

[38] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Mor-

ishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 11 2016.

[39] X. D. Zhang, J. Song, P. Bork, and X. M. Zhao. The exploration of network motifs as potential drug targets from post-translational regulatory networks. *Sci Rep*, 6:20558, Feb 2016.

[40] F. Cheng, I. A. Kov?cs, and A. L. Barab?si. Network-based prediction of drug combinations. *Nat Commun*, 10(1):1197, 03 2019.

[41] Y. Hou, J. Zhao, W. Martin, A. Kallianpur, M. K. Chung, L. Jehi, N. Sharifi, S. Erzurum, C. Eng, and F. Cheng. New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. *BMC Med*, 18(1):216, 07 2020.