UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Master Degree in Physics of Data

Final Dissertation

Characterizing SARS-CoV-2 oncogenic features from protein-protein multilayer interaction networks

Thesis supervisor

Prof./Dr. Manlio De Domenico

Thesis co-supervisor

Prof./Dr. Vera Pancaldi

Candidate

Francesco Zambelli

Academic Year 2022/2023

Thanks to the NetBIO team in Toulouse for their warm welcome and for making me feel at home. They have not only treated me as a colleague but also as a friend, someone worth spending extra time with.

Thanks to Professor Manlio De Domenico for introducing me to this field and having great confidence in me. His trust made me feel like I was contributing to something important.

Thanks to my friends back home for for being my roots. With some of them I feel like I could engage my entire life.

To my friends in Padova, being my first and most important achievement in my independent life. In particular Gaetano and Giacomo, with whom I shared a year of living together that now feels like a beautiful miracle.

Lastly, thanks to my family for being my nest, a place I can always return to, but also I place I'll have to leave.

Abstract

In recent years, network-based approaches have become increasingly popular in the field of molecular biology and network medicine. One of the main challenges in this field is to identify the role of viruses in the development of cancer. In this thesis, we apply multilayer complex network theory to protein-protein interaction networks of different viruses, including the SARS-CoV-2 virus, to classify them as oncogenic or non-oncogenic. Specifically, we use tools from graph theory and machine learning to analyze the topology and structure of these networks, and to identify the key proteins and pathways involved in virus-induced carcinogenesis. We aim to create two classes, one for oncogenic and one for non-oncogenic viruses, and then verify in which of the two the SARS-CoV-2 virus falls. The results of this study may provide new insights into the mechanisms underlying virus-induced cancer and could lead to the development of novel therapeutic strategies.

Chapter 1

Research question

As a consequence of the global Covid-19 pandemic, the importance of understanding the intricate relationship between viruses and cancer has gained unprecedented significance. The increasing significance that this dimension will have on the people's life, it has become increasingly crucial to anticipate and prepare for the potential long-term consequences, with the objective to be able to face them on time. A special role is given to the relation between viral infections and cancer, one of the most lethal illnesses worldwide. Exploring the interplay between viruses and cancer not only sheds light on the mechanisms of disease, but also offers valuable insights into forecasting the potential impact of infections on future patients with cancer.

The research surrounding virus-cancer interactions takes on added urgency against the backdrop of the pandemic. Viruses, such as the SARS-CoV-2 virus responsible for COVID-19, can have profound effects on human health beyond the acute phase of infection. Emerging evidence suggests that certain viral infections, including COVID-19, may have implications for cancer development and progression in the long term[1–5]. Understanding the molecular and cellular mechanisms by which viruses influence the oncogenic process is therefore essential for forecasting the potential rise in cancer cases in the aftermath of the pandemic.

1.1 Tumor insurgence and the role of viral infections

Cancer is a complex and multifaceted disease that involves the uncontrolled growth and spread of abnormal cells in the body. Normally, cells in the body divide and grow in a regulated manner, replacing damaged or dying cells and maintaining healthy tissues and organs. However, cancer cells break free from this normal growth control mechanism and start to divide and grow uncontrollably, forming a mass of cells called a tumor. These cells can invade nearby tissues and organs, and even spread to other parts of the body through the bloodstream or lymphatic system, a process known as metastasis.

Cancer can arise from different types of cells in the body, and there are more than 100 different types of cancer, each with its own set of characteristics, behavior, and treatment options. Some of the most common types of cancer include lung, breast, prostate, colon, and skin cancer.

Within a tumor, the cells are surrounded by a variety of immune cells, fibroblasts, molecules and bold vessels. These components create the system called tumor microenvironment. The tumor and the surrounding microenvironment are closely related and interact constantly. Tumors can influence the microenvironment by releasing extracellular signals, promoting tumor angiogenesis and inducing peripheral immune tolerance, while the immune cells in the microenvironment can affect the growth and evolution of cancerous cells. The role of the surrounding immune cells is also very complex: they could be of different types, T cells, B cells, macrofages, and, depending on the kind of tumor and other condistions, they can be either pro-tumorigenic or anti-tumorigenic.

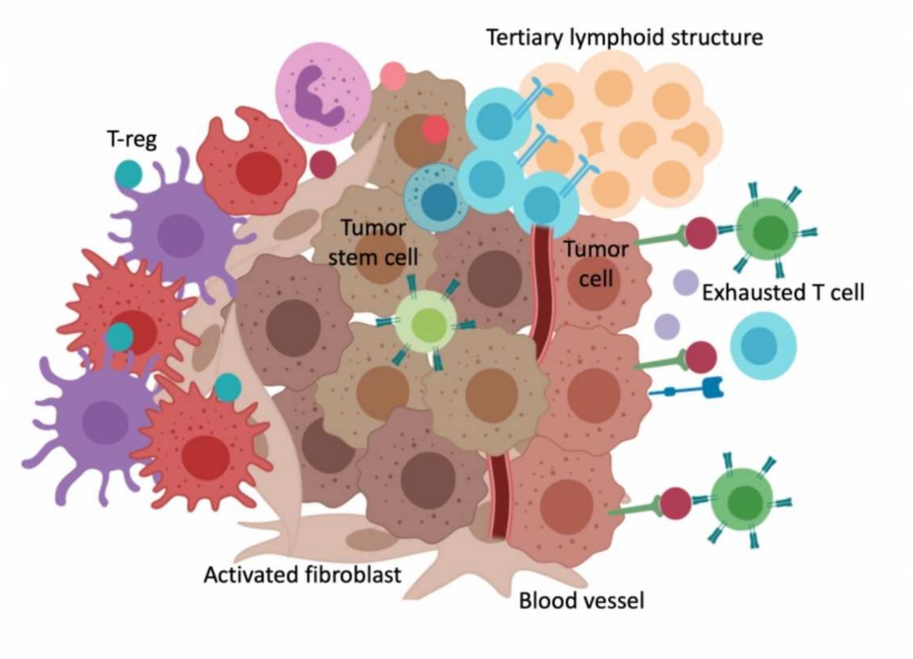


Figura 1.1: A graphical schematic representation of the tumor micro-environment and some of its key components. The figure is taken from *The tumor microenvironment. (2020, May 17). Keren Lab.* `https://www.weizmann.ac.il/mcb/Keren/research-activities/tumor-microenvironment`

The complexity of the tumor microenvironment and its interaction between its components, complicates even further the understanding of cancer and differentiate a lot the causes, the symptoms and the possible cures between different cases.

The exact causes of cancer are not fully understood, but it is generally believed to be a result of a combination of genetic and environmental factors. Certain genetic mutations or alterations in the DNA of cells can increase the risk of cancer development, while environmental factors such as exposure to carcinogens, radiation, or viral infections can also contribute to the onset of cancer.

Cancer cells exhibit several characteristics that differentiate them from normal cells:

- They are capable of growing without the need for signals that promote growth, while normal cells require signals to grow and divide.
- They often ignore signals that normally control cell division and programmed cell death (apoptosis).
- They can invade nearby tissues and metastasize to distant organs in the body. Normal cells do not typically move throughout the body and stop growing when they contact other cells.
- They are able to promote the growth of new blood vessels towards the tumor (angiogenesis), which helps to supply the tumor with nutrients and oxygen and remove waste products.
- They can evade the immune system's normal function of detecting and destroying abnormal or damaged cells.
- They may even manipulate the immune system to help the tumor cells grow and survive.
- They accumulate genetic changes, such as duplications or deletions of chromosomes, at a higher rate than normal cells. Some cancer cells even have double the usual number of chromosomes.
- They often rely on different nutrients and energy sources compared to normal cells, allowing them to grow and divide more quickly.

As the scientific research about cancer developed, an increasing number of infections have been linked to the development of specific human cancers. It has been outlined that slightly more than 20% of the

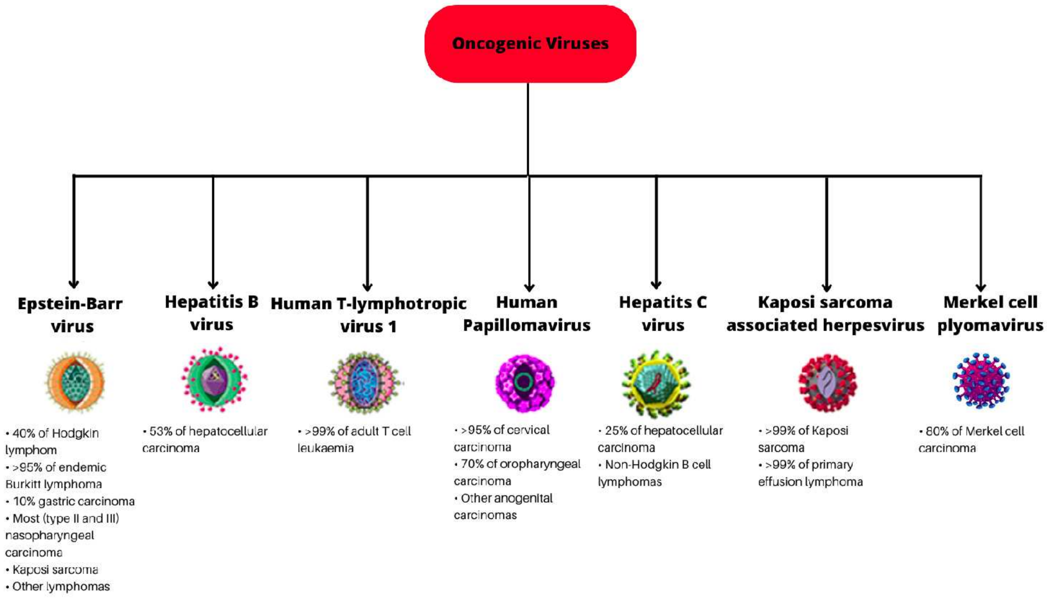

Oncogenic Viruses
Epstein-Barr virus
Hepatitis B virus
Human T-lymphotropic virus 1
Human Papillomavirus
Hepatits C virus
Kaposi sarcoma associated herpesvirus
Merkel cell plyomavirus
• 40% of Hodgkin lymphom
• >95% of endemic Burkitt lymphoma
• 10% gastric carcinoma
• Most (type II and III) nasopharyngeal carcinoma
• Kaposi sarcoma
• Other lymphomas
• 53% of hepatocellular carcinoma
• >99% of adult T cell leukaemia
• >95% of cervical carcinoma
• 70% of oropharyngeal carcinoma
• Other anogenital carcinomas
• 25% of hepatocellular carcinoma
• Non-Hodgkin B cell lymphomas
• >99% of Kaposi sarcoma
• >99% of primary effusion lymphoma
• 80% of Merkel cell carcinoma

Figura 1.2: The following is a compilation of tumor-associated viruses, with information on the incidence of each tumor type sourced from [6].

global cancer incidence arises as the consequence of previous infections, often caused by the presence of a virus[6]. In these cases, continued expression of specific viral genes seems to be a prerequisite for maintaining the malignant phenotype of the arising cancers.

Viruses have generally a very small genome in which most genes encode proteins involved in virus replication, while only a very small portion of the genes may be related to tumor cell transformation. Although different viruses encode different virus products, they can target some of the same mechanisms, such as inhibiting tumor suppressor gene expression, abnormally activating oncogenes to interfere with cell growth and differentiation-related signals, thereby affecting cell growth cycle regulation and inducing malignant cell transformation[7].

Such viruses, that can cause cancer in humans and animals, are called oncogenic viruses. These viruses are able to insert their genetic material into the host cell's DNA, which can lead to the activation of oncogenes or the inactivation of tumor suppressor genes. Oncogenes are particular sections of the DNA which encodes information for the production of proteins which have a very important role for the promotion of the tumor (in this case they will be promoted), or for the suppression of it (and in this case they will be inhibit). Some examples of such genes are:

- HER2/neu (also known as ErbB2): is involved in the regulation of cell growth and differentiation, and its overexpression is seen in several types of cancer, including breast, ovarian, and gastric cancer.
- BCL-2: is involved in the regulation of apoptosis, or programmed cell death, and its overexpression can lead to the survival of cancer cells that would normally die.
- MYC: regulates cell proliferation and is often overexpressed in various types of cancer, including breast, lung, and colon cancer.
- RAS: family of oncogenes that regulates cell signaling pathways that control cell growth and division, and mutations in RAS genes are found in many types of cancer, including pancreatic, lung, and colorectal cancer.
- TP53 (also known as p53): This tumor suppressor gene helps to prevent the formation of cancer by regulating cell growth and apoptosis. Mutations in TP53 are found in a wide range of cancers, including lung, breast, and colorectal cancer[8, 9].

- pRB: is a proto-oncogenic tumor suppressor protein that is dysfunctional in several major cancers. One function of pRb is to prevent excessive cell growth by inhibiting cell cycle progression until a cell is ready to divide[10, 11].

The cellular tumor antigen refers to the TP53 and retinoblastoma protein (pRB), which are central to the two main tumor suppressor pathways. These pathways regulate cell cycle progression, stimulate DNA damage response, and induce apoptosis in response to irreversible cell damage. While nearly all oncogenic viruses encode oncoproteins that disrupt the TP53 and pRB pathways, the mechanisms they use are distinct. Viral oncoproteins inhibit p53 and pRB by inducing their degradation, inactivation, repression, or dissociation from functional partners.

In this work this problem is tackled by exploiting methods form complex network theory. Firstly the virsu-host cell interaction are modelled as a PPI network, in which the nodes are human genes influenced by the action of the virus, and the links are the structural, functional, physical connections between them. After doing this it's possible to apply over such systems the tools from network theory, with the goal to find out specific feature that can allow to identify oncogenic viruses, with the final objective to state if Sars-Cov2 could belong to this category or not.

Finally it's important to mention also oncolytic viruses, a type of virus that can selectively infect and kill cancer cells while sparing healthy cells. They work by infecting cancer cells and replicating within them, causing the cells to burst and die. This releases new virus particles that can infect neighboring cancer cells, leading to the destruction of the tumor. In addition to directly killing cancer cells, oncolytic viruses can also stimulate an immune response against the tumor. When cancer cells are destroyed by the virus, they release molecules that can activate immune cells such as T cells and natural killer cells to recognize and attack remaining cancer cells. This immune response can help to eliminate any remaining cancer cells and prevent the tumor from returning.

1.1.1 Sars-Cov2

SARS-CoV-2 is a novel coronavirus, responsible for the COVID-19 pandemic. It is a single-stranded RNA virus that belongs to the family Coronaviridae, which also includes the viruses SARS and MERS [12, 13]. The SARS-CoV-2 virus is spherical in shape and has a diameter of approximately 120 nm. The outer surface of the virus is covered in spike proteins, which give it a "crown" or "corona" appearance under the microscope. These spike proteins are responsible for binding to and entering host cells, which they do by binding to the ACE2 receptor on the surface of human cells.

Once the virus enters the host cell, it releases its RNA genome, which is then translated into viral proteins by the host cell's machinery. These proteins include replicase proteins, which are responsible for replicating the viral RNA, and structural proteins, which are used to build new virus particles.

The symptoms of COVID-19 can range from mild to severe, and can include fever, cough, shortness of breath, fatigue, muscle or body aches, and loss of taste or smell. In severe cases, the virus can lead to pneumonia, acute respiratory distress syndrome, and death. Due to the fact that the outbreak and the virus itself are pretty recent, it's likely that many other symptoms and consequences on the organism after the infection, are still to be discovered.

A novel approach comprehending the use of multilayer networks involving virus-host, host-host, gene-symptoms, diseases-illness, diseases-symptoms and gene-drugs connections[14] proposes a method to investigate the relations between the biological characteristics of the virus and the infection consequences on the host, and was able to both confirm already observed features and predict some others.

The global impact of the virus and its widespread infection among the population have sparked a surge in research efforts. The need to comprehensively understand the virus's effects on the human body, particularly its long-term implications, has become paramount. This has resulted in an ongoing stream of studies dedicated to investigating various aspects of the virus and its impact on health.

Such research is crucial for developing effective strategies to mitigate the long-term consequences and address the challenges posed by this global pandemic.

Chapter 2

Network Medicine

In recent years, the study of complex networks has undergone significant advancements, which have expanded its influence to various fields. One of the major breakthroughs has been the application of complex network theory to medical science, giving birth to network medicine, an interdisciplinary field that combines the principles of network science, systems biology, and medicine to study complex diseases and their underlying mechanisms[15–17]. This approach focuses on understanding how the components of biological systems, such as genes, proteins, and metabolites, interact with each other in a network fashion to maintain health or contribute to disease. By modeling and analyzing these networks, network medicine aims to identify key molecular players and pathways that can be targeted for therapeutic intervention[16, 18, 19].

Network medicine has already shown promising results in the study of a variety of diseases, including cancer, neurological disorders, and infectious diseases. For example, it has been used to identify novel drug targets and biomarkers, and to predict drug responses and disease outcomes[14, 20]. Moreover, network medicine approaches have the potential to transform clinical practice by enabling personalized medicine, where treatments are tailored to individual patients based on their unique molecular profiles and network characteristics.

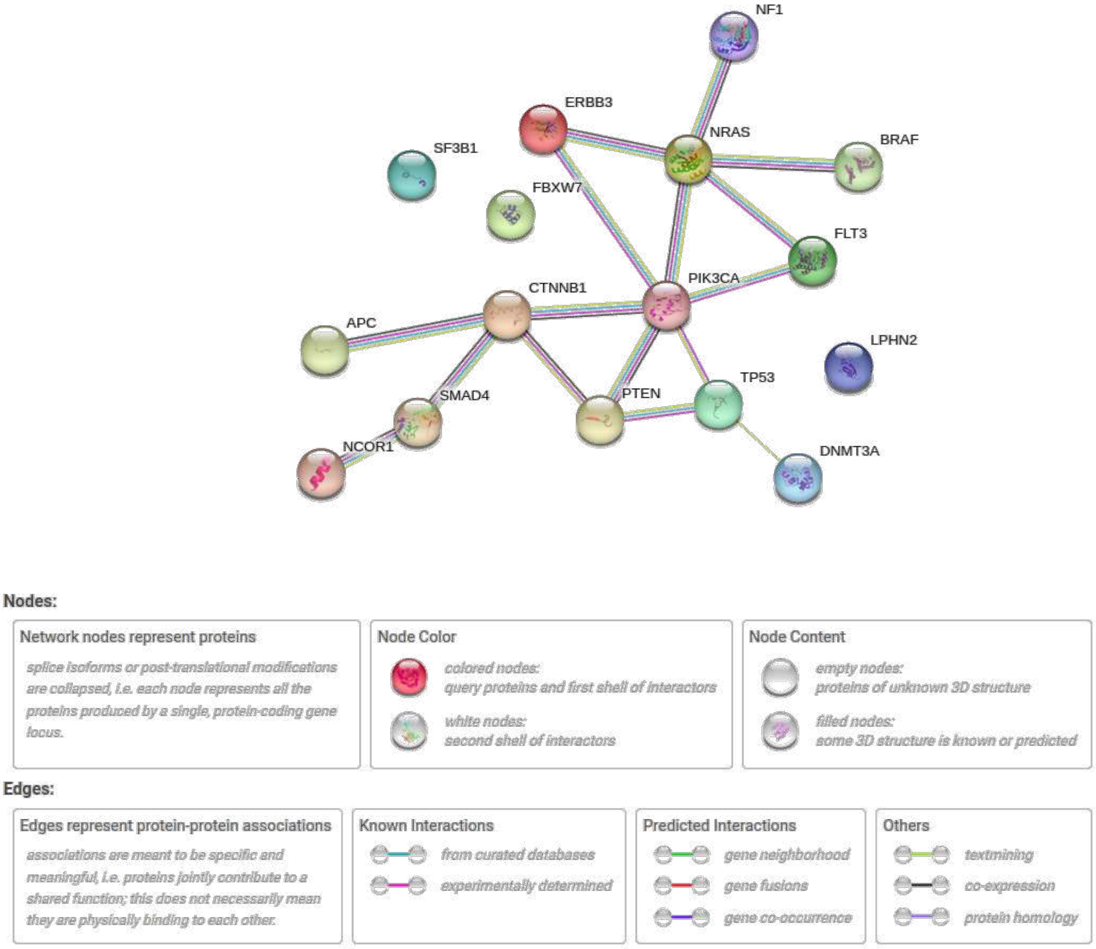
Overall, network medicine represents a powerful framework for understanding complex biological systems and developing new strategies for disease prevention, diagnosis, and treatment.

To further leverage the potential of complex network theory, the concept of multilayer networks or multiplex systems has been introduced. Multilayer networks are composed of several replicas of the same nodes, each connected in a unique way, with each replica connected to some of its other replicas. This approach enables the consideration of various aspects of a system simultaneously and being able to study them in a global and integrated way. A new set of mathematical tools has been developed to study such complex systems, as detailed in the paper [21].

2.1 Network Medicine and PPI networks

A complex network is essentially composed by two entities: nodes, which can correspond to physical cosntituents of the system under consideration, and edges, which represent the connections presented between the nodes. Both nodes and edges can widely vary in nature and type, but in evry case it's possible to build from a physicl system a mathematical model that can be studied in datail tanks to a set of computational anf theoretical tools (such aspects will be further explained and deepened in chapter 2.4).

A complex network consists of nodes and edges, representing the constituents and connections of a system, respectively. In this work, we focus on protein-protein interaction (PPI) networks, which describe biological systems or processes. Proteins are represented as nodes, connected by edges that



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes: query proteins and first shell of interactors

white nodes: second shell of interactors

Node Content

empty nodes: proteins of unknown 3D structure

filled nodes: some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

Figura 2.3: Example of graphical representation of a PPI interaction network involving the 16 most frequently mutated human cancer genes from the STRING web page `https://string-db.org/cgi/input?sessionId=bnKXMWMsf0q5&input_page_active_form=examples`. The legend depicts the different nature of the links between the proteins and some information about the nodes.

represent various types of interactions. These interactions can range from physical interactions observed in experiments to correlations in gene expression or even co-citations in scientific articles. Such heterogeneity is discussed in detail in Chapter 2.4.

Fig. 2.3 illustrates a PPI network depicting interactions between the 16 most frequently mutated human cancer genes in STRING. The edges in this network can be categorized as known interactions (curated databases or experimental evidence), predicted interactions based on gene relationships, or inferred using text mining or structural similarity. It is important to note that the edges in PPI networks capture multiple aspects of the original systems rather than a homogeneous physical quantity.

Complex network theory provides computational and theoretical tools to analyze PPI networks and extract valuable information. These tools, described in detail in Chapter 2.4, enable the assessment of node importance, network resilience to attacks or failures, identification of network components, and detection of meaningful communities or clusters of proteins.

By analyzing these network properties, we can gain insights into the underlying biological system. For example, identifying important nodes in the network can help identify potential therapeutic targets, while understanding network resilience can provide insights into disease susceptibility. Additionally, the identification of protein communities can reveal functional modules or pathways within the system.

Overall, complex network analysis of PPI networks offers a powerful approach to studying biological systems and addressing specific biological questions.

PPI networks have limitations in capturing the full complexity of biological systems, as they often overlook the interplay of multiple aspects. However, the emergence of multilayer networks in network medicine has addressed this limitation by integrating diverse biological dimensions into a comprehensive framework. This approach has gained significant momentum in recent years, yielding valuable insights and significant advancements in understanding complex biological contexts. By considering multiple layers of information simultaneously, multilayer networks offer a more holistic perspective, enabling researchers to unravel intricate relationships and uncover novel findings with far-reaching implications in the field of network medicine [14, 22, 23].

2.2 PPI networks for virus-host interaction study

PPI networks reveals to be a good tool also to study the interaction between two organism which interacts at a cellular level, and in particular virus-human interactions [24–28]. Viruses interact with human cells at the protein level through a variety of mechanisms. When a virus infects a host cell, it aims to hijack the cellular machinery and utilize it for its own replication and propagation. The interactions between viral proteins and human proteins play a crucial role in this process. One common interaction is the binding of viral surface proteins to specific receptors on the surface of human cells. This initial attachment allows the virus to gain entry into the host cell. Once inside the cell, viral proteins interact with various cellular proteins to manipulate the host cell's processes and create an environment conducive to viral replication. Viral proteins may interact with host cell proteins involved in gene expression, protein synthesis, immune response, and cellular signaling pathways. These interactions can modulate host cell functions to favor viral replication and evade the immune system.

The interaction between viruses and human cells can be effectively analyzed within the framework of protein-protein interactions (PPIs). Various models can be employed to describe these interactions. For instance, viruses can modify existing protein connections or disrupt them altogether. They may also establish new connections between regions that were not previously linked. In this study, the focus is on investigating the impact of viral infections on the human cells by subsetting the entire human PPI down to the specific region likely to be influenced by the virus, as outlined in detail in Chapter 3.2.3.

2.3 Why Multilayer Networks?

Multilayer networks can be used in network medicine to provide a more comprehensive and realistic representation of complex biological systems [14, 22, 23, 29]. Traditional network models often consider interactions between entities as a single layer, disregarding the potential heterogeneity and multiple dimensions of these interactions. However, in many biological systems, interactions occur across different layers or modalities, such as genetic, protein-protein, and metabolic interactions, or even considering at the same time different states in which the same system can be, and so being able to explore the relation between these.

By incorporating multilayer network analysis in network medicine, researchers can capture the intricate interplay between different layers of biological information and gain a deeper understanding of disease mechanisms. Multilayer networks allow for the integration of diverse data types, such as genomics, proteomics, and clinical data, into a unified framework. This integration enables the exploration of complex relationships and dependencies between various biological components, offering a more holistic view of disease processes.

The use of multilayer networks in network medicine provides several advantages. First, it allows for the identification of key nodes or molecules that play critical roles across multiple layers, highlighting their significance in disease progression or therapeutic targets. Second, it enables the investigation of cross-layer interactions and their impact on disease outcomes. By considering the interconnectedness between different layers, researchers can uncover hidden associations and uncover novel biomarkers or pathways that may not be apparent in individual network layers. Third, multilayer networks facilitate the development of personalized medicine approaches by capturing patient-specific molecular profiles and incorporating them into a multilayer framework for tailored treatment strategies.

Network medicine offers a compelling approach for cancer research, recognizing the complexity and heterogeneity of this disease. Cancer exhibits diverse manifestations and employs various strategies to affect the body. In this context, network medicine emerges as an ideal framework to gain insights into the intricate systems underlying cancer. By employing specialized tools and methodologies, network medicine allows researchers to explore the complex interactions among genes, proteins, and other molecular components involved in cancer. This integrative approach facilitates a comprehensive understanding of the disease, enabling the identification of crucial players, pathways, and interactions that drive cancer progression. Thus, network medicine holds great promise for advancing our knowledge of cancer and paving the way for improved strategies in its prevention, diagnosis, and treatment.

2.4 Enrichment Analysis

By leveraging the capabilities of network medicine, valuable insights can be extracted from the studied systems. In this work, the analysis is focused on protein-protein interaction (PPI) networks, where key features often involve specific lists of proteins. To gain biological insights related to the underlying research question, functional enrichment analysis plays a crucial role, and becomes one of the most important tool to connect the mathematical framework to the biology of the system.

Functional enrichment analysis is a widely used bioinformatics approach that aims to understand the biological meaning behind a set of genes or proteins by identifying overrepresented functional categories or pathways. It helps researchers gain insights into the underlying biological processes, molecular functions, or cellular components associated with a particular gene or protein set.

The analysis typically involves the following steps:

1. Data Preparation: The gene set of interest is identified and collected. This could involve selecting genes based on statistical criteria or prior knowledge.

2. Background Selection: A background set, often representing all genes or proteins in a reference genome, is defined to establish a baseline for comparison.

3. Statistical Analysis: Various statistical methods, such as hypergeometric test, Fisher's exact test, or chi-square test, are applied to assess the significance of functional enrichment. These tests evaluate whether the observed number of genes in a particular functional category is significantly higher than what would be expected by chance, given the background set.

4. Functional Database: A curated database, such as Gene Ontology (GO), KEGG, Reactome, or other pathway databases, is used to provide functional annotations for genes or proteins. These databases categorize genes based on their biological functions, molecular processes, or involvement in specific pathways.

5. Multiple Testing Correction: Since functional enrichment analysis involves testing multiple functional categories simultaneously, multiple testing correction methods, such as Bonferroni correction or false discovery rate (FDR) adjustment, are applied to control for the inflation of false positives.

6. Result Interpretation: Enrichment analysis results are typically presented as a list of enriched functional categories or pathways, along with statistical measures such as p-values or FDR-adjusted p-values. Researchers interpret these results to gain insights into the biological processes, molecular functions, or pathways that may be associated with the gene set of interest.

There are many possible choices of functional database to use in the functional enrichment analysis, each of them characterizing a different aspect of the system under study, and depending on the context and the biological question to be answered, during the following work, many of these will be used.

The principal online tool that is utilized is *ToppGene* [1], a comprehensive web-based platform that provides a range of bioinformatics tools and resources for gene function prediction, gene prioritization, and functional enrichment analysis.

[1] `https://toppgene.cchmc.org/`

Chapter 3

Methods from Complex Network Theory

Complex network theory is a subfield of physics that has recently experienced significant progress. Its primary objective is to develop mathematical frameworks that encode information about the interconnections between various parts of a system. These frameworks are then used to extract valuable information from the system. The complexity of these mathematical frameworks can be increased through the use of multi-layer network formalisms. Such formalisms allow for the consideration of a system in different contexts but at the meantime it's able to preserve the distinction between them.

To classify the risk associated with the SARS-CoV-2 virus, it is necessary to identify relevant features and compare them to distinguish between oncogenic and non-oncogenic viruses. Complex networks and multi-layer complex networks are tools used for this purpose. The following paragraph provides a brief introduction to these topics and presents tools for feature extraction, categorized into three main groups: microscale, mesoscale, and macroscale methods.

3.1 Complex networks, monoplex

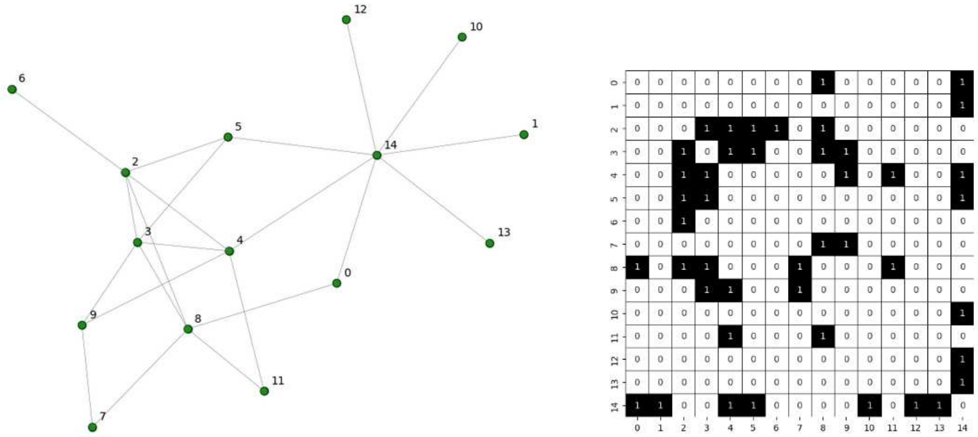
Complex networks, also known as complex graphs, are systems composed of nodes, or vertices, that are interconnected by links, or edges. The most basic forms of complex networks are simple graphs, which consist of at most one edge between a pair of nodes and no self-edges, meaning connections from a vertex to itself. Another critical aspect that distinguishes networks, particularly based on the nature of the edges, is the differentiation between directed and undirected networks. In directed networks, the edges have directionality, meaning they point from one vertex to another. In contrast, undirected networks lack directionality, and a link only represents the presence of a connection between two nodes.

To increase the complexity of the mathematical representation of complex networks, specific weights can be added to each edge, representing the strength of that particular link. Such networks are called weighted networks.

3.1.1 Adjacency Matrix

The adjacency matrix is the most well-known and significant mathematical representation of a complex network. It is a matrix A with dimensions nxn, where n represents the number of nodes in the network. Each entry in the matrix, $A(i, j)$, indicates if there is a connection between node i and node j and also encodes the strength of such a link (Fig.3.4).

Symmetric adjacency matrices are used to represent undirected networks. If a link between nodes i and j is present, the link between node j and i will also exist, with the same weight. Hence, the spectral theorem holds in such cases, and we can be certain of finding positive and real eigenvalues.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0
3	0	0	1	0	1	1	0	0	1	1	0	0	0	0	0
4	0	0	1	1	0	0	0	0	0	1	0	1	0	0	1
5	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
8	1	0	1	1	0	0	0	1	0	0	0	1	0	0	0
9	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	1	1	0	0	1	1	0	0	0	0	1	0	1	1	0

Figura 3.4: On the left a simple 15 nodes undirected unweighted network. On the right its representation as an adjacency matrix.

In the simplest formulation, where the network only indicates if a link between two nodes is present or not without giving information about their relative strengths, the adjacency matrix is built as follows:

$$A_{i,j} = \begin{cases} 1 & \text{if there is a link between node i and node j,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

In contrast, adjacency matrices of directed networks are not necessarily symmetric. Therefore, when computing characteristic quantities like eigenvalues and eigenvectors, greater care is required [30].

3.1.2 Microscale Methods

Microscale methods are tools used to analyze the structure of a network by examining individual nodes and links to obtain information about the network's overall structure or specific components within it.

Degree and Degree Distribution

The degree of a node refers to the number of edges connected to it. In directed networks, in-degree is the number of links arriving at a node, while out-degree is the number of links leaving it. For undirected networks with N nodes, the degree of a node can be determined using the adjacency matrix:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{3.2}$$

Connectance or density can be calculated from the degree by determining the fraction of actual edges in the network compared to the total number of possible edges. The total number of possible edges is equal to $\binom{N}{2}$ for a network with N nodes and M edges. The equation for connectance is:

$$\rho = \frac{M}{\binom{N}{2}} \tag{3.3}$$

The degree distribution $P(k)$ is the fraction of nodes in a network with a given degree k. If there are N nodes in total and n_k nodes have a degree of k, then $P(k) = \frac{n_k}{n}$. Understanding the degree distribution of complex networks is a crucial aspect of network science. Two main branches of thought have emerged: one suggests that degree distributions follow a power law, implying that complex networks exhibit scale-free properties. This conclusion may be reached because in continuously growing

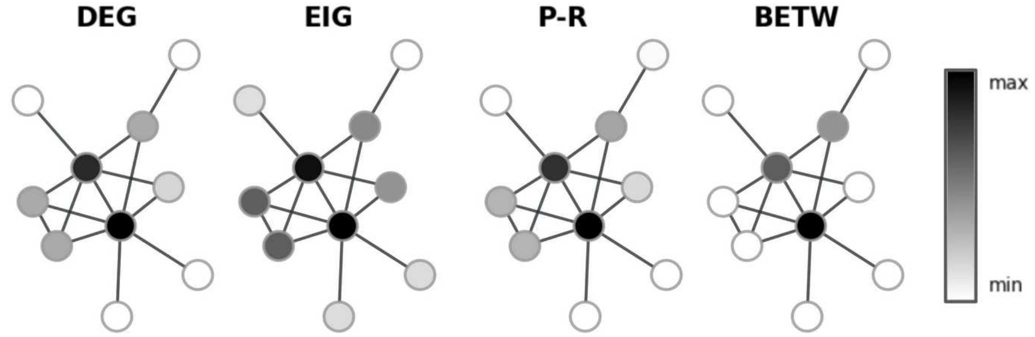


Figura 3.5: Illustrative examples of centrality measures on a simple network consisting of 10 nodes. Each centrality measure is visually represented by the color of the nodes, ranging from white for the least important nodes to black for the most central ones. From left to right, the centrality measures depicted are: Degree centrality, Eigenvector centrality, PageRank centrality, and Betweenness centrality.

networks, new nodes are likely to connect to nodes that are already well connected. To check for this property, one can look at the log-log plot of the histogram of the degree distribution to determine whether it follows a straight line. In many cases, the degree distribution takes on an exponential-like shape in both the small and large k regimes, leading to debate about whether this effect is solely due to finite-size effects in real-world networks or if it has a deeper meaning. [31, 32]

Measuring Node Importance: Centrality

When analyzing a network, a crucial question that arises is which nodes are the most significant in the system. However, the concept of importance can be ambiguous, and there are various ways to interpret it. Therefore, multiple methods have been developed to measure node importance, or centrality. A graphical representation of some examples of centrality assignment for a very simple network can be found in Fig.3.5.

Degree Centrality

The simplest method to measure node importance is degree centrality, which looks at how many connections a node has. Each node is assigned a measure proportional to its degree, which is the number of nodes it is connected to. Although it is a basic measure, it can be illuminating, as nodes with many connections are likely to play significant roles in the network.

Eigenvector centrality

Eigenvector centrality is a measure of the importance of a node in a network based on the idea that the importance of a node increases if it is connected to another node that is itself important. In contrast to degree centrality, which only looks at the number of connections a node has, eigenvector centrality takes into account the characteristics of the nodes that are connected to it.

To calculate eigenvector centrality, we use an iterative process. We start by assigning an initial score of 1 to each node. We then update the vector of scores to a new vector by summing the scores of the neighbors of each node. This can be written mathematically as:

$$x'i = \sum j = 1^N A_{ij} x_j \tag{3.4}$$

In vector form, this can be written as $\mathbf{x}' = \mathbf{A}\mathbf{x}$. After t iterations, we have:

$$\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0) \tag{3.5}$$

It is possible to write any initial score vector $\mathbf{x}(0)$ as a linear combination of the eigenvectors of $\mathbf{A}$, i.e., $\mathbf{x}(0) = \sum_i c_i v_i$. Using this, we can rewrite the previous expression as:

$$\mathbf{x}(t) = \mathbf{A}^t \sum_i c_i v_i \tag{3.6}$$

This iterative process allows us to calculate the eigenvector centrality of each node in the network, taking into account the importance of the nodes it is connected to.

Random walks and PageRank centrality

PageRank is a measure of centrality originally developed to rank web pages[33]. The centrality value of each node is proportional to the probability that a random walker, subject to teleportation, will occupy that node while traversing the network. The probability of transitioning from one vertex to another is proportional to the strength of the edge connecting them.

To calculate the probability of transitioning from node i to node j, we use the formula: $T_{i,j} = \frac{A_{i,j}}{d_i}$, where A is the adjacency matrix and d_i is the out-degree of node i. This ensures that the sum of probabilities from node i to all other possible nodes is equal to 1.

By defining the matrix D as the diagonal matrix with the values of the out-degree d_i, we can define the transition matrix as:

$$T = D^{-1}A \tag{3.7}$$

To find the occupation probability, we define the vector $p(t)$ as the vector whose i_{th} entry corresponds to the probability for the walker to be at node i at time t. The evolution of this vector is given by:

$$p(t+1) = p(t)T \tag{3.8}$$

The occupation probability corresponds to the steady state of this process, given by:

$$p^*(t) = p^*(t)\mathbf{T} \tag{3.9}$$

This is equivalent to finding the leading eigenvector of the transition matrix $\mathbf{T}$, which corresponds to the eigenvalue 1. Therefore, p^* is equal to the leading eigenvector of $\mathbf{T}$.

Ultimately, the vector p^* will give the asymptotic probabilities of the random walker to be in the different nodes of the network given the transition probabilities defined with T.

Percolation

Percolation in the context of complex networks refers to the phenomenon where the connectivity or information flow in a network undergoes a sudden and significant change as nodes or edges are gradually removed.

The key idea in percolation is to study the emergence of a connected component that spans the entire network, known as the largest connected component (LCC). The LCC represents the largest cluster of connected nodes that allows for efficient flow or transmission of information, influence, or other phenomena across the network.

In the context of complex networks, percolation can help understand various phenomena such as the spread of diseases, the robustness of infrastructure networks, or the efficiency of information flow in social networks. By analyzing percolation properties, researchers can gain insights into the critical thresholds or conditions at which the network undergoes a significant change in connectivity or function. Such quantity correspond to a certain fraction of nodes (or edges) removed from the network in which a transition form a ordered to disordered phase occurs in the system. In order to spot it, the best way it so look for the maximum of the second largest component size depending on the fraction of nodes (or edges) removed.

3.1.3 Mesoscale methods

Mesoscale methods in network theory refer to approaches and techniques that analyze the intermediate level of organization in complex networks, which lies between the microscopic level of individual nodes and edges, and the macroscopic level of the entire network. These methods aim to uncover patterns, structures, and properties that emerge at the mesoscale, providing insights into the collective behavior and functional modules within complex networks.

Components

When analyzing a network, it's very important so see how an information can spread though the system. For this reason it's useful to study the components, which are defined as groups of nodes in the network in which every pair of origin and target are connected by at least one path, and it's not possible to add any other vertex of the network while preserving such a property. For networks with finite size we define the *Largest connected component (LCC)* as the cluster with the maximal subset of nodes. If there exist a path from every vertex in the networks to every other, the network is connected; otherwise we talk about disconnected network.

Bayesian Communities Detection (DCSBM)

Communities detection is a fundamental task in complex network theory, which aims at identifying groups of nodes with high density of connections within the group and low density of connections between groups. In other words, communities are groups of nodes that are more densely connected to each other than to the rest of the network. The detection of communities is useful for many applications, such as understanding the structure and function of networks, identifying functional modules in biological systems, detecting groups of users with similar interests in social networks, and so on. There are various methods for community detection, which can be broadly classified into two categories: divisive methods that iteratively split the network into smaller sub-networks, such as the Girvan-Newman algorithm [34], and agglomerative methods that iteratively merge nodes or sub-networks into larger communities, such as the famous Louvain algorithm [35]. These methods can be based on different criteria, such as optimizing modularity, maximizing likelihood, minimizing conductance or cut, or using spectral or probabilistic methods. The choice of the method and criteria depends on the specific characteristics of the network and the application at hand.

In this section we will analyse specifically the Stochastic Block Model (SBM) and it variant in which a correction on the degree distribution of nodes in each group is applied, the Degree Corrected SBM (DCSBM). A more complete discussion of the argument can be found in [36]. The SBM assumes that the system can be thought as a realization of an underlying model encoding some mechanism that generate what can be observed. The aim is to find a set of parameters of the generative model that maximize the posterior probability $P(\Theta, \mathbf{b}|A_{i,j})$ given by the product of a prior distribution of the parameters $P(\Theta, \mathbf{b})$ to be decided in advance, and the likelihood of the data, which corresponds to the structure of the system encoded by the adjacency matrix $A_{i,j}$ given a certain set of parameters $P(A_{i,j}|\Theta, \mathbf{b})$.

$$P(\Theta, \mathbf{b}|A_{i,j}) = \frac{P(A_{i,j}|\Theta, \mathbf{b})P(\Theta, \mathbf{b})}{P(A_{i,j})} \tag{3.10}$$

The parameters of the systems are $\mathbf{b}$, a vector of length equal to the number of nodes in the network n, which describe the partition of nodes in B groups, where $B \in [0, n-1]$, and Θ, a list of additional model parameters that control how the node partition affects the structure of the network.

The SBM then specifies a probability distribution for the edges in the network based on the communities to which the nodes belong. Specifically, it assumes that the probability of an edge between two nodes depends only on the communities to which they belong. The probability of an edge between nodes in the same community is typically higher than the probability of an edge between nodes in different communities.

Although quite general, the traditional SBM model assumes that the edges are placed randomly inside each group, and because of this the nodes that belong to the same group tend to have very similar degrees. As it turns out, this is often a poor model for many networks, which possess highly heterogeneous degree distributions. A better model for such networks is called the degree-corrected stochastic block model [37], and it is defined just like the traditional model, with the addition of the degree sequence $\mathbf{k}$ of the graph as an additional set of parameters.

The SBM is a powerful method for analyzing the structure of complex networks with communities. It can be used to detect communities in networks, to compare the community structure of different networks, and to generate synthetic networks with known community structure for testing algorithms and models.

3.2 Multilayer networks

Complex systems often operate in multiple contexts, and the agents of the system interact differently in each context. Multilayer networks theory provides a mathematical framework for representing multiple representations of the same system. In a multilayer network, the same set of nodes represents different contexts in which the elements corresponding to the nodes can operate. For instance, in network medicine, a system may be composed of proteins, genes, and metabolites that interact in different contexts, such as metabolic processes, responses to particular external stimuli, and gene regulation. Each of these contexts describes a different layer of the same multilayer network representation of a cell.

In order to address the requirements of this study, a dedicated computational toolset was developed. The toolset, named *MuxVizPy*, was created by modifying the existing R library *MuxViz* [38] and introducing new functionalities to overcome computational challenges and incorporate additional tools. This adaptation allowed for the efficient analysis and visualization of multilayer networks, tailored to the specific needs of the research.

Multilayer networks are represented by a set of nodes and a set of edges, but with some extra peculiarities compared to the monoplex case. When discussing nodes, we can consider the entity encoding a certain element of the system, which is called the physical node, and is represented in each layer of the network. Alternatively, we can think of a specific replica of such a node in a particular layer, which is called a state node.

One major difference between multilayer networks and monoplex networks is the presence of edges that can be of different types, which can be divided into two main categories. The first category is intralayer interactions, which refer to interactions between nodes in the same layer. These interactions can be further classified as self-interactions ($\mathbb{S}$), where the edge is from a node to itself, or endogenous interactions ($\mathbb{N}$), which are links between distinct nodes in the same layer. The second category is interlayer interactions, which include exogenous interactions ($\mathbb{X}$), which are links between nodes belonging to different layers, and intertwining interactions ($\mathbb{I}$), which are links from a node to its replica in other layers.

We can classify different types of multilayer structures based on the presence of edges of certain categories (a graphical representation is shown in Fig.3.6):

1. Multiplex interconnected: This type of multilayer structure consists of networks that have edges of types ($\mathbb{S}$, $\mathbb{N}$, $\mathbb{I}$). In this case, the interlayer connections consist only of edges between replica nodes.

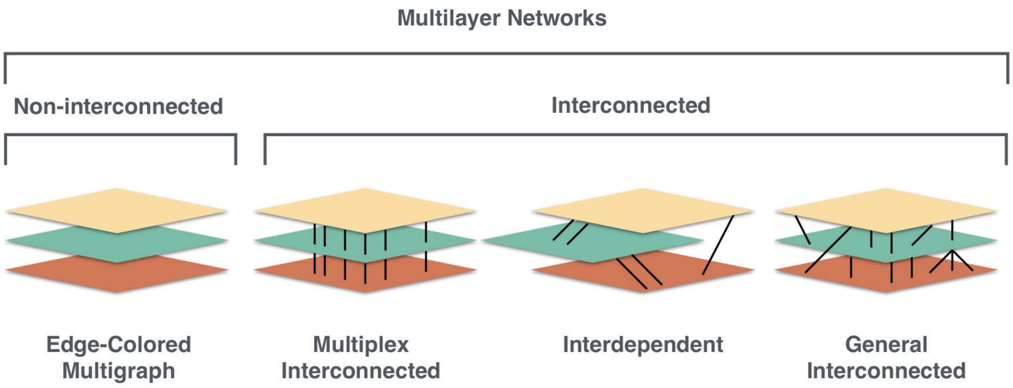


Figura 3.6: Basic classification of multilayer network models based on their interlayer links. Figure from [39] under Creative Commons AttributionShareAlike 4.0 International License.

2. Interdependent networks: In this type of multilayer structure, we have edges of types ($\mathbb{S}$, $\mathbb{N}$, $\mathbb{X}$), and replica nodes are not connected to each other.
3. General multilayers: This type of multilayer structure can have connections of all types, including edges of types ($\mathbb{S}$, $\mathbb{N}$, $\mathbb{X}$, $\mathbb{I}$).

On the other hand, when the interconnection between layers is not well specified, it's possible to use the edge-colored representation, in which transitions between replicas of a same node are uniform and constant.

3.2.1 Representation of Multilayer Networks

Due to the higher dimensionality structure of multilayer networks, new tools had to be developed to efficiently represent them. This section introduces the tensorial representation[21], which is a formalism that enables the production of rigorous theoretical results, and the supra-adjacency matrix, which is a flattened representation of the multi-adjacency tensor. The supra-adjacency matrix is often used when dealing with the computational part because it enables the use of all the mathematical tools developed for 2-dimensional matrices.

The tensorial representation of a multilayer network allows to specify whether a node i in a given layer α is connected to another node j in layer β. To facilitate mathematical operations such as outer and inner products, the tensor is built using the covariant and contravariant formalism. We define a tensor $M^{j_1,j_1,\ldots,j_m}_{j_1,j_1,\ldots,j_n}$ of rank mn, m-covariant, and n-contravariant. With this notation, we can define the two main operations. The outer product of two tensors X and Y is a new tensor Z with a number of covariant (contravariant) indices equal to the sum of the number of covariant (contravariant) indices of X and Y. Therefore, the outer product of two tensors is always a tensor of higher order than the original ones. For example, $X^{k}_{ij}Y^{mn}_{l} = Z^{kmn}_{ijl}$.

It is also possible to define an inner product, which corresponds to a contraction because the rank of the resulting tensor is reduced by two units. For instance, this is the case in the product $X^{k}_{ij}Y^{mn}_{l} = Z^{mn}_{ij}$, where the index k is covariant for X and contravariant for Y. This operation corresponds to summing over the components of X and Y identified by the index k.

Using these mathematical tools, it is possible to define a multilinear object in the space $\mathbb{R}^{NxLxLxN}$, which corresponds to the multi-adjacency tensor:

$$M^{i\alpha}_{j\beta} = \sum_{a,b=1}^{N}\sum_{\alpha,\beta=1}^{L} w_{ab}(pq)e_i(a)e_j(b)e_\alpha(p)e^\beta(q) = \sum_{a,b=1}^{N}\sum_{\alpha,\beta=1}^{L} w_{ab}(pq)E^{i\alpha}_{j\beta}(ab;pq) \tag{3.11}$$

where $e_i(a)$ and $e^j(b)$ are the covariant and contravariant canonical rank-1 tensors, and $w_{ab}(pq)$ encodes the intensity of the interactions between node a in layer p and node b in layer q.

It's important to mention that we focus on multilayer networks with interlayer connectivity since it is not possible to define a meaningful multilayer adjacency tensor for edge-colored multigraphs.

Although working with tensors may be difficult and cumbersome from an operational perspective, it provides a very compact and useful way to write complex equations and can guide our intuition for generalizing network descriptors for multilayer analysis. Matricization or flattening is a widely adopted approach that maps a high-order tensor into a lower-order object while preserving the information content. In the case of a multilayer adjacency tensor, flattening results in a rank-2 object known as a supra-adjacency matrix. Although working with a supra-adjacency matrix has computational advantages, it also has notational disadvantages. The supra-adjacency matrix is a block matrix with intralayer connectivity encoded in diagonal blocks and interlayer connectivity encoded in off-diagonal ones. However, this arrangement of blocks is not unique, and other arrangements are also valid. This non-uniqueness makes the supra-adjacency matrix less suitable for theoretical calculations but still a viable alternative to higher-order tensors.

3.2.2 Microscale methods

The topological tool presented in the section before to investigate the features that the single nodes plays in the network are here further developed and adapted to the multilayer network framework.

Multi-Degree

Muilti-degree centrality is the simplest indicator of a node importance at a local level, it is obtained by summing up all the links connected to node i across all layers.

$$k_i = \sum_{\alpha,\beta=1}^{L} \sum_{j=1}^{N} M^{i\alpha}_{j\beta} \tag{3.12}$$

If the interconnection between layers are not well defined, it's more suitable to sum up all the degree contribution of each layer separately:

$$k_i = \sum_{\alpha=1}^{L} k_i^{\alpha} = \sum_{\alpha=1}^{L} \sum_{j=1}^{N} A_{i,j}(\alpha) \tag{3.13}$$

where A^i_j is the adjacency matrix of the layer α.

Multi-Eigenvector Centrality

The ratio behind the definition of the eigenvector centrality in the multilayer scenario is the same of the monoplex case. When the intralayer connections are known (i.e. we are not considering the edge-colored case), the natural extension of the leading eigenvector problem that allows to compute the eigenvector centrality in the single layer scenario, consists in setting up the leading eigentensor problem for the multi-adjacency tensor:

$$\sum_{i,\alpha} M^{i,\alpha}_{j,\beta} \Theta_{i\alpha} = \lambda_1 \Theta_{j\beta} \tag{3.14}$$

with λ_1 the leading eigenvalue of M and Θ the leading eigentensor. The final goal of a multilayer-centrality measure is to have a score for each physical node in the network, and, given the fact that Θ encodes information about all the replica nodes, it's necessary to define a procedure to summarize all the information about each replica node of a physical node in a single value. They possible choices are multiple, but usually the most common one is to sum up all the contributions:

$$\theta_i = \sum_{\alpha} \Theta_{i\alpha} \tag{3.15}$$

where θ is the final vector containing the values of the multi-eigenvector centrality for each physical node.

Computationally speaking, as reported in the previous sections, it's more convenient to work with the supra-adjacency matrix rather than the multi-adjacency tensor. In this framework the problem stated before resolves in finding the leading eigenvector of the supra adjacency matrix. It's then necessary to keep in memory which position in the diagonal of the supra-adjacency matrix represent which replica node, and them merge the entries corresponding to the same physical node with the chosen criteria, as example by summing them up.

When dealing with very big networks (in this work the number of replica nodes of a multilayer network are of the order of 10k), the computational power needed for both storing in memory the supra-adjacency matrix but also for solving the leading eigenvector problem, becomes very high. To solve the problem two solutions were adopted and implemented in `MuxVizPy`. For the lack of memory, rather than dense matrices storing the values of each entry, spares matrices were adopted, supposing that the number of actual connections between nodes if far smaller than the total number of possible links ($\binom{N}{2}$). For the leading eigenvector problem, an iterative approximation procedure was adopted. We start from a vector with arbitrary entries x_0, which is multiplied at each time-step with the matrix in consideration producing a new vector x_{new}. If the euclidean distance between x_0 and x_{new} is smaller than a given threshold, the algorithm stops, otherwise we set $x_0 = x_{new}$ and iterate the process. The algorithm stops either when it reaches convergence or when a maximum number of steps are done. Such an approximation is acceptable when considering the ultimate goal to be reached, especially when it reaches convergence with a small value of the tolerance parameter (10^{-8} in the actual library implementation), which rather than obtaining exact values for each entry, consists in obtaining an order of importance of the nodes, and to be able to differentiate different order of magnitude of the "importance" values.

Multi-Random Walks, Transition Matrix and Multi-Page Rank versatility

Like in the single layer scenario, the multi-pagerank centrality measure whats to find a score for each node corresponding to the occupation probability of that node by a random walker in the network. Setting up the random walker dynamic in the network corresponds in finding the multi-transition tensor $T^{i\alpha}_{j\beta}$, which gives for each node in each layer the transition probability in any other node in an other layer. Usually such a quantity is computed as proportional to the weight of the corresponding edge, and the probabilities of a transition from a given node to all the other nodes in the network, should be normalized and sum up to 1. With the multi-transition tensor it's possible to obtain the occupation probabilities for each replica node by solving the leading eigentensor:

$$T^{i\alpha}_{j\beta}\Pi_{i\alpha} = \lambda_1 \Pi_{j\beta} \tag{3.16}$$

The probabilities $\Pi_{i\alpha}$ defines the *random walk occupation centrality*, which can be further compressed by aggregating values corresponding to replicas of the same node obtaining a vector of scores for each physical node.

On the other hand *multi-pagerank centrality* can be obtained by solving the same eigentensor problems, but rather than considering the multi-transition tensor $T^{i\alpha}_{j\beta}$, we suppose that the random walker is subjected to teleportation: at any time the random walker can walk to a random neighbor with a rate r, and be teleport to any other node with rate $1 - r$. The new transition tensor which is created in this way is thus:

$$R^{i\alpha}_{j\beta} = rT^{i\alpha}_{j\beta} + \frac{1-r}{NL}u^{i\alpha}_{j\beta} \tag{3.17}$$

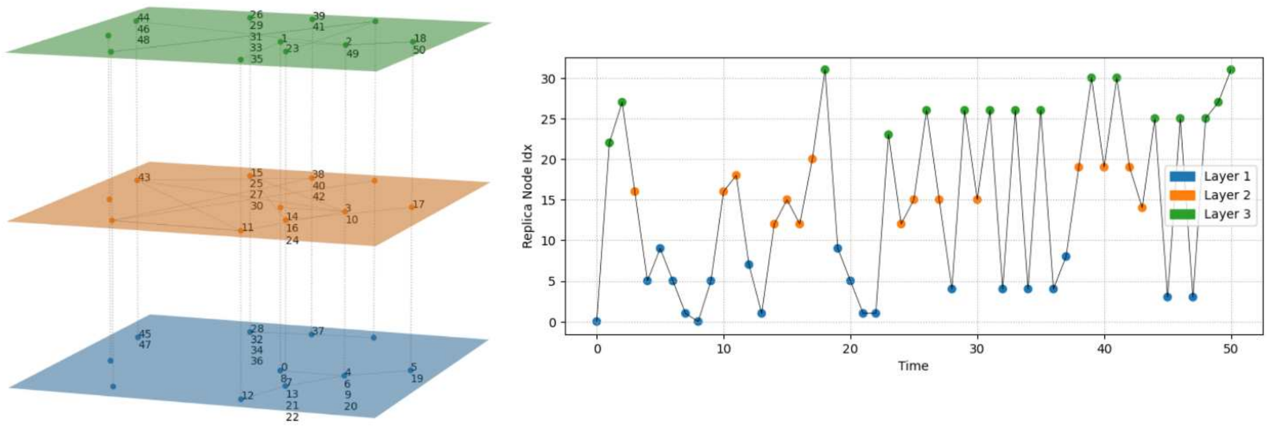


Figura 3.7: The figure illustrates a 50-step random walk on a multiplex interconnected network. On the right, the plot shows the replica node where the random walker is located at each timestep. The replica node index is calculated as $physical_node\_idx * layer\_idx$. On the left, the representation of the multilayer network is displayed, with the numbers near the nodes indicating the timestep when the random walker traverses them.

with N the number of physical nodes, L the number of layers and $u^{i\alpha}_{j\beta}$ the rank-4 tensor with all the entries equal to 1.

Similar to the case of the supra-adjacency matrix, also the multi-transition tensor can be ”flattened” in a standard matrix of dimension $NLxNL$, which is called supra-transition matrix. In this way the leading eigentensor problem can be mapped in a standard leading eigenvector problem. The computational issues are the same of the case of the eigenvector centrality, and the solutions adopted in `MuxVizPy` are also the same for the *random walk occupation centrality.* For the *multi-pagerank centrality* summing up a constant matrix to the supra-transition matrix (which is reasonable to suppose as sparse), doesn't allow anymore to use the sparse matrix tool. The solution becomes then to not compute directly $R^{i\alpha}_{j\beta}$, but adding its contribution at each iteration of the leading eigenvector approximation algorithm described in the previous section. In this way it's also possible to compute the *multi-pagerank centrality* in a very accurate and efficient way.

It's worth highlighting that the result of both the *multi-pagerank centrality* and *random walk occupation centrality* are higly dependent on the values of the interaction strengths of the interlayer links. In this work, only multiplex interconnected are considered, so all the replicas of the same physical nodes are connected with each other. It's important to tune properly the strengths of this interactions, as example they can be related to the difference of the degree distribution of the layers. If all the links of the multiplex have the same weight, and in one of such layer the mean degree of the nodes is much larger than the number of layers in the multiplex, the probability of a transition between different layers is much smaller than the one of remaining in the same layer. This could bias the final result, by leading to bad performances.

To mitigate these challenges, a decision was made to treat the multilayer network as an edge-colored graph during the application of the page-rank algorithm. Instead of explicitly defining interlayer links, a uniform probability of transition between layers was considered for a random walker. This approach helps to circumvent biases that may arise from poorly tuned interlayer link strengths and leads to more robust and reliable results.

Percolation

In the context of percolation, there are two types to consider in the single-layer case: node percolation and edge percolation. However, in the multilayer case, the options expand. In terms of edges, one can remove individual edges from the single layers or eliminate all edges between replica nodes corresponding to the same physical nodes. Similarly, for nodes, one can progressively remove all connections related to a replica node or remove all connections from all replica nodes corresponding to the same physical node.

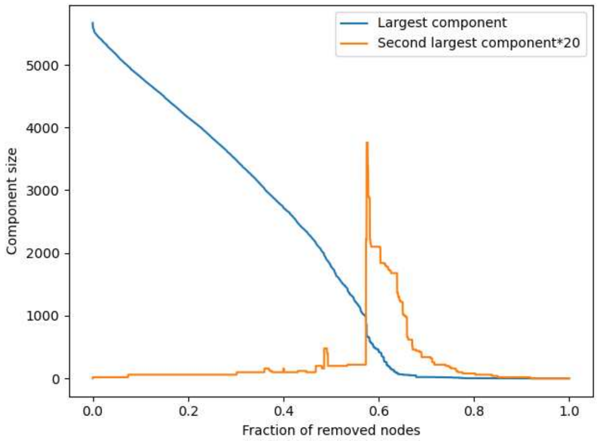


Figura 3.8: Example of a physical node percolation result for a interconnected 7 layers multilayer network produced with the PPI networks data. From the trends of both the first and second component it's clearly visible the transition between an ordered and a disordered state of the system, correlated to the dimension of the largest components.

Each type of percolation can be carried out in different ways, depending on the order in which nodes or edges are removed. This can be done randomly to simulate random failures in the system, or it can follow a specific metric order to simulate an external attack on the system.

Once a procedure for network disintegration is established, it is important to calculate certain topological features as the process unfolds to monitor the changes in the system. The most common choice is to track the Largest Connected Component (LCC), but other measures such as the Largest Interconnected Component (LIC) or the Viable Component (LVC) will be discussed in subsequent paragraphs.

When considering the LCC, a critical point becomes a crucial related quantity. It represents the fraction of nodes or edges that need to be removed in order to transition from an ordered state, where the LCC size is greater than 0, to a state where the system loses connectivity and becomes completely scattered. This critical point is also associated with the peak of the second largest component size. An example of such a process can be observed in Fig.3.8.

3.2.3 Mesoscale methods

Components

The analysis of the components is the analyzed aspect in which there is the bigger difference between the single layer and multilayer scenarios. In fact when looking at the "fundamental" components of a multiplex, the quantities to be take into consideration are 3:

- **LCC**(largest connected component): is the components containing the maximum subset of nodes. It means that there isn't any other node in the network which is connected by at least one multilayer path with one of the nodes in the LCC.
- **LIC**(largest intersected component): it corresponds to the set of nodes that are connected in all the layers independently.
- **LVC**(largest viable component): consists in the set of nodes which are connected by the same path in all the layers independently. Such nodes are usually very important, because they are crossed by paths which are present simultaneously in all the nodes, and for this reason it's very likely that are fundamental in the overall system behaviour.

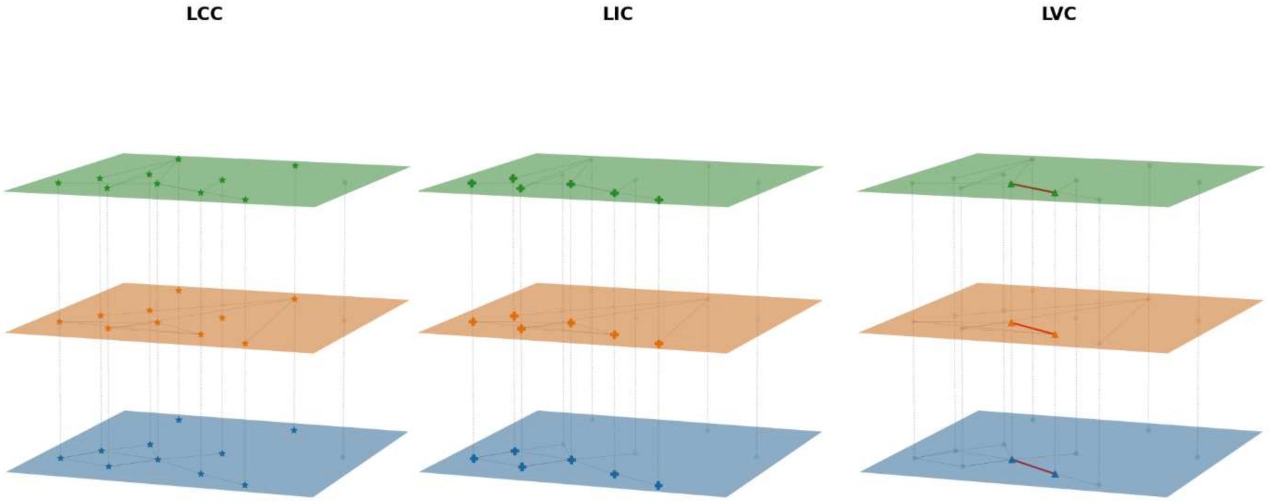


Figura 3.9: LCC, LIC and LVC graphical representation for a simple multiplex interconnected. In the LVC it is highlighted with a thick red line the edge shared between all the layers.

DCSBM applied to edge-colored networks

When dealing with the multilayer networks framework, also an extension of the DCSBM algorithm presented previously is available[40]. It works with a layered structure, which correspond to the edge-colored multilayer netowrk case, in which the interlayer connections are not specified. In this case the DCSBM takes into account both the node degrees (number of connections) and the community assignments across all layers to generate a probabilistic model of the network. The algorithm considers a degree correction factor for each node and layer, which accounts for the variation in node degrees across layers. This correction factor allows for the modeling of nodes that may have different degrees in different layers. By incorporating this degree correction, the algorithm can capture the inherent heterogeneity in node connectivity across layers.

The DCSBM algorithm in layered structures involves estimating the model parameters, such as the community assignments and the degree correction factors, using likelihood maximization or Bayesian inference techniques. These parameters are optimized to best fit the observed network data across all layers. The resulting model can then be used to analyze the network's community structure, identify influential nodes or communities, and understand the interplay between different layers in the network.

Overall, by extending the DCSBM to layered structures, we can gain insights into the complex interactions and relationships within multi-layer networks, providing a more comprehensive understanding of the system under study.

Chapter 4

Dataset

In this work, the focus is on analyzing interactions between viruses and human proteins using complex network analysis, specifically utilizing a multilayer network structure. Protein-protein interaction networks, which are complex graphs connecting nodes representing proteins with various types of relationships, play a crucial role in this framework.

Obtaining data on virus-host interactions from the literature involves multiple sources, each studying different viruses and employing diverse methods to gather interaction data[41]. Assembling a comprehensive database that encompasses all the information from the literature is beyond the scope of this work. Therefore, multiple approaches were proposed and their differences were thoroughly analyzed.

A particular focus will be given to data from the STRING database [42], a bioinformatics resource that provides information on protein-protein interactions (PPIs). It serves as a comprehensive repository of known and predicted protein interactions, consolidating data from various sources such as experimental studies, computational predictions, and text mining. STRING assigns confidence scores to protein interactions, indicating the reliability of the evidence supporting each interaction.

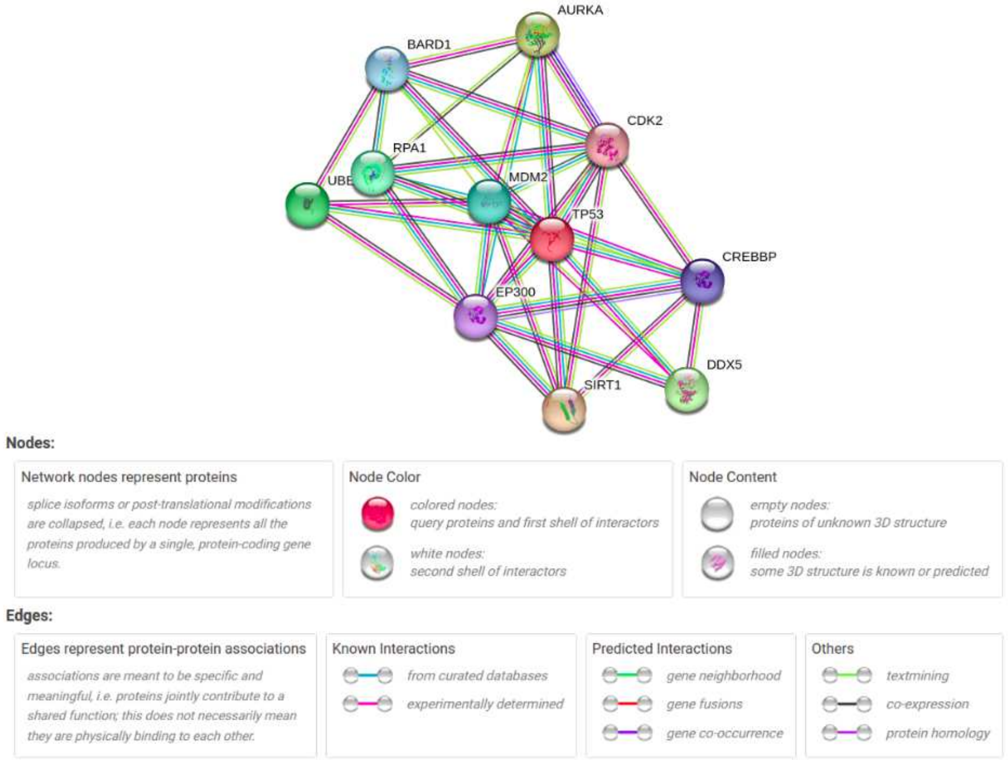
Starting from the STRING dataset, appropriate virus-host protein-protein interaction networks can be constructed to form the dataset for this work. The choice of network construction method depends on the initial assumptions made for the analysis, as different approaches can provide unique insights into the problem at hand. By comparing and identifying common results among these approaches, the aim is to strengthen the conclusions of this work, avoiding dataset-specific biases and enhancing the generalizability of the findings.

Another important resource is viruses.STRING, a online database created from STRING data which focuses on the description between the interactions between viruses and host[43]. Essentially it properly select the subset of the host PPI which interacts with the virus protein, building itself PPI network which can be either downloaded as a file and even analyzed graphically.

4.1 PPI links experimental search

Protein-protein interaction databases can contain different types of links that represent different types of interactions between the proteins. Here are some examples of the types of links that can be found in STRING and viruses.STRING:

1. **Physical interactions:** These links represent direct physical interactions between two proteins, such as binding, catalysis, or transport. There are many ways to investigate such links in laboratory. As example it's possible to put markers in different organisms (e.g. yields), merge them together and see if such marked proteins interact, or by using particular techniques to move a specific protein, and looking if the other one follows it. Another way it to see how an organism responds to the removal or one of the two or both the considered proteins, exploiting



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes: query proteins and first shell of interactors

white nodes: second shell of interactors

Node Content

empty nodes: proteins of unknown 3D structure

filled nodes: some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

Figura 4.10: Example of graphical representation of a PPI interaction network involving the links related to the TP53 protein from the STRING web page `https://string-db.org/cgi/input?sessionId=bMApwDqLCTiS`

particular techniques such as CRISPR: if it survives when one of the two is removed, but dies when both of them are eliminated, it means that probably they are strongly related for some important biological process.

2. **Functional interactions:** These links represent indirect interactions between two proteins, such as proteins that participate in the same biological pathway, share a functional domain, or regulate each other's activity.

3. **Co-expression links:** These links are based on gene expression data and represent proteins that are co-expressed across multiple experiments, which may suggest that they are functionally related.

4. **Text mining links:** These links are based on automatic text mining of scientific literature and represent proteins that are mentioned together in the same article, which may suggest a functional association.

5. **Gene neighborhood links:** These links represent proteins that are produced by genes close to each other in the network, and may suggest a functional relationship even if there is no direct evidence of interaction.

6. **Gene fusion links:** These links include proteins that are produced by genes that in a certain point of the evolution, merged together. Also in this case there is no evidence of actual interactions, but the connection is considered to be likely and inferred.

7. **Genes omology links:** They represent connections between proteins that are produced by genes which present a similar spatial conformation.

The protein-protein interaction (PPI) network of viruses in STRING database incorporates various types of links from diverse sources and prediction methods, providing a comprehensive view of protein interactions involved in viral infections.

Each link is assigned a confidence score, determined through different approaches based on the method used to identify the link.

To construct a reliable dataset for analysis, it is important to implement a meaningful slicing procedure that retains only the links consistent with the initial assumptions for constructing a specific type of PPI network. However, it should be acknowledged that due to the nature of link determination procedures, this approach may not capture the complete biological context. In the following paragraph, we discuss the implications of this incompleteness.

4.2 PPI network incompleteness and afterwards

As we delve into our research, it is crucial to keep in mind that the protein-protein interaction networks we are dealing with are incomplete. When an edge exists in the dataset, it indicates a probable connection between the genes, but the degree of confidence in its presence can vary. To address this, we must select a threshold and only consider edges with a likelihood of being present above that value. The higher the score, the higher the confidence in the existence of such links and a the lower is the probability of false positives.

However, false negatives are more common, as detecting a link requires conducting an experiment to verify it. With networks as large as the human proteome, which has a vast number of possible links, it is impossible to verify all of them. Additionally, the number of links related to a specific protein can reflect its popularity, as more experiments are likely to be conducted on important proteins, leading to more interactions being discovered.

Taking all of these factors into account, and selecting interaction with high scores, we will assume that our networks are complete enough to provide useful insights in our subsequent analyses.

4.3 From Databases to Networks

The methodology employed in this study to construct functional protein-protein interaction (PPI) networks that accurately represent the interaction between human cells and viruses is inspired by the approach proposed by the CoMuNeLab [44]. The required data consist of two types: firstly, information describing the human PPI network, which includes a list of links between pairs of human proteins selected based on specific criteria, enabling the construction of the human PPI network.

Additionally, data pertaining to how the virus interacts with human proteins are needed to appropriately represent this interaction in network form. Specifically, the data should describe the interactions between human and viral proteins, enabling the identification of human proteins directly targeted by the virus.

Following this step, two approaches can be pursued. The first approach involves subsetting the entire human PPI network by considering only the nodes and edges related to the human proteins directly targeted by the viruses. However, with the available data, this approach often yields small and disconnected networks.

To address this limitation and create more connected networks, an alternative approach involves including other proteins that are likely to be influenced by the virus. The most straightforward choice is to select the nearest neighbors of the directly targeted proteins and extract the corresponding subset from the entire human PPI network. This approach, depicted in Fig.4.11, is preferred as it allows for better utilization of network theory tools and will be utilized to construct the networks for the subsequent analyses.

The resulting networks, constructed using this approach, provide insight into how viral infections impact the human proteome. Different networks corresponding to different viruses can be considered as descriptions of the same system (the human proteome) under the influence of distinct external stimuli (viral infections).

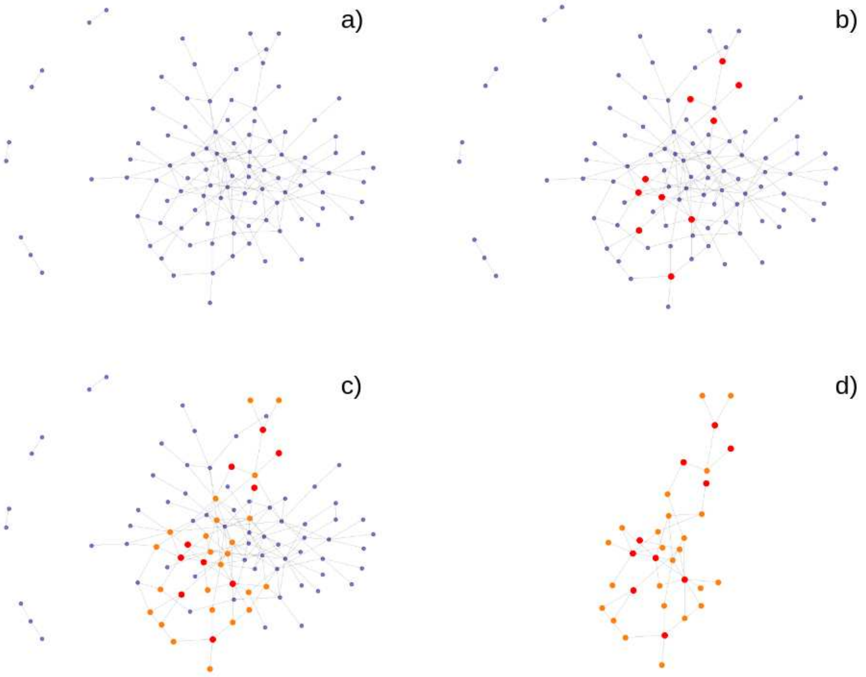


Figura 4.11: The figure illustrates by using a simple 50 nodes random network, how the method to build the PPI networks associated to each virus works. Firstly the entire human PPI network is considered (a). Then the nodes which are directly targeted by the virus are considered (b), and in order to enlarge the network, also their nearest neighbours are picked (c). Finally only the subset of the entire original network involving the nodes highlighted in the previous steps is extracted resulting in the final PPI network describing the virus-host interaction.

4.4 BioSTRING

The first dataset used is the one proposed by the CoMuNeLab[44], called BioSTRING, a proteomen interaction database build by integrating the PPI databases obtained both form the database BioGRID[45] v3.5.18217,33 (publicly available at `https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.182/`), while the second one is STRING[42] v11.016 functional interactions network (publicly available at `https://stringdb.org/cgi/download.pl`). To avoid nomenclature problems, the data from both dataset were filtered by using the NCBI gene database to map all protein names and aliases to a common nomenclature of official symbols. Specifically, were used the data made publicly available from NCBI at `ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/`(Accessed March 28, 2020).

The nature of the links between different proteins obtained both by BioGRID and STRING can be very different, they can be related to functional connections, co-occourence in some systems, physical interactions, physical similarities, but all these aspects are taken together and a final aggregate network describing the interaction between human proteins is built.

The majority of the data containing the interactions between the viruses proteins and the host-human proteins are taken from STRING Viruses[43], an online database derived from STRING in which it's possible to obtain data of virus-host PPI networks for a great variety of both virus and host species. From this source were obtained PPI interactions for 80 viruses which can infect the humans. As regard the human-human proteins interactions, BioSTRING provides 737668 links which involve 19946 different proteins, that are used to buid the entire human PPI network.

By starting form this network, the procedure presented before allows to create 80 host-virus PPI networks describing how the human proteome is influenced by the infection of the virus. An example of such a network is reported in Fig.4.12.

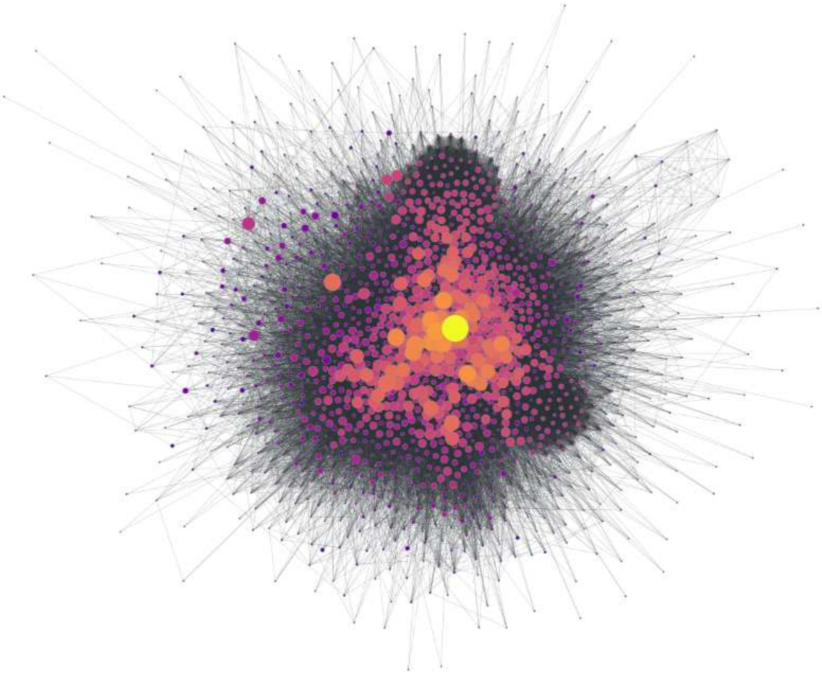


Figura 4.12: A graphical representation of the PPI nearest neighbor network associated with the *Human rhinovirus A serotype 89*. The node size in the network corresponds to the betweenness centrality, while the color of the nodes indicates the corresponding page rank centrality. Nodes with low page rank centrality are depicted in blue, while nodes with high page rank centrality are depicted in yellow.

At the end for each virus are generated two files. The first one contains a list of nodes and their types, which are categorized as viral genes (0), directly targeted human genes (1), higher-order human genes (2), disease (3), symptom (4), drug (5), or biological process (6). The types of nodes from 3 to 6 are integrated by exploiting other interaction resources, which will not be used and presented, so, for the purpose of this work, only nodes of type 1 and 2 were considered.

The second file contain a list of the edges between pairs of nodes contained in the nodes file. These edges can involve nodes of all six different types, but due to our focus on nodes of types 1 and 2, only edges between these nodes were considered.

In order to be able to build PPI networks as representative of the real system as possible, these links obtained from viruses.STRING, representing the interaction between the host and the virus, are all characterized by high confidence level, which mean grater than 0.7.

4.4.1 Viruses in BioSTRING

The dataset includes information for each virus, such as the virus name, the number of proteins it contains, the number of interactions with human proteins, the virus family, and a crucial attribute for this study: a boolean value indicating whether the virus is categorized as oncogenic or not. Based on this information, the viruses can be divided into two groups: oncogenic and non-oncogenic.

The oncogenic group consists of the following eight viruses: Epstein-Barr virus strain B95-8, Hepatitis B virus genotype C subtype ayr isolate Human-Japan-Okamoto, Hepatitis C virus genotype 1a isolate H, Human herpesvirus 8 type P isolate GK18, Human papillomavirus type 16, Human papillomavirus type 18, Human papillomavirus type 5, and Human T-cell leukemia virus 1 isolate Caribbea HS-35 subtype A.

The non-oncogenic group comprises 72 viruses. However, when considering only the non-oncogenic viruses, the SarsCov2 virus, which is the target of this analysis, is excluded. Therefore, the effective number of non-oncogenic viruses is 71.

4.4.2 Problems in BioSTRING

During the analysis of the BioSTRING database, some critical issues were identified. Firstly, the database was found to be based on an outdated version of the protein-protein interaction (PPI) data from STRING. This raises concerns about the accuracy and relevance of the interactions included in the database.

Furthermore, upon examining the raw data, it became apparent that the majority of the considered links, those with a confidence level greater than 0.7, were attributed to text-mining-based interactions. This poses several drawbacks for further analysis. The number of edges associated with a particular protein is likely to be influenced by the number of publications on that protein, which may not necessarily correspond to a higher number of actual connections. This aspect is not a feature of interest for the current work. Additionally, it was observed that experimentally determined interactions typically have medium confidence scores ranging from 0.45 to 0.7, rather than high confidence levels (> 0.7). Therefore, by solely selecting high confidence links, it is possible to retain interactions that are deemed highly probable but assign minimal importance to experimentally derived links.

To address these concerns, a decision was made to construct a new dataset following a similar approach to BioSTRING. However, in this new dataset, these considerations will be taken into account both during the construction of the human protein-protein interaction network and for the identification of interactions between viruses and human proteins. By doing so, a more accurate and reliable dataset can be generated, facilitating the subsequent analysis.

4.5 "Co-expression" and "Experimental" Databases

To address the issues mentioned earlier, a new dataset was created based on the STRING database.

Initially, a dataset containing all interactions involving Homo sapiens was downloaded. This dataset includes interactions among human proteins as well as interactions with proteins from other organisms, such as viruses.

The next step involved constructing the human protein-protein interaction (PPI) network. To do this, a set of predefined criteria was used to filter the interactions from the dataset. The aim was to avoid biases introduced by literature-based sources. Specifically, only links supported by gene co-expression were chosen, indicating a statistically significant correlation between the expression of the two proteins. The confidence level threshold for these links was set to 0.4, ensuring a medium level of confidence.

To expand the dataset while still excluding literature-based sources, additional links discovered through other experimental procedures were considered. Two additional sources from the STRING database, namely "experiments" and "co-occurrence," were incorporated. "Experiments" encompassed various detection methods, such as Yeast Two-Hybrid, Co-immunoprecipitation, Affinity Chromatography, and Mass Spectrometry, representing links investigated through laboratory experiments. "Co-occurrence" referred to the observation that two proteins tend to interact or coexist in the same cellular context more frequently than expected by chance, suggesting a functional or regulatory relationship.

Table 4.1 presents some topological features of each of these human PPI networks. As expected, each of the three filtered networks contained less than 10% of the total number of interactions in the original database, reflecting the exclusion of low-confidence links. The co-expression network was approximately three times smaller than the experimental network and more than six times smaller than the high-confidence network, with a similar trend observed for the number of nodes.

Protein nomenclature poses another aspect to consider. The existence of multiple naming systems for proteins introduces confusion and inconsistency when referring to the same protein across different studies or databases. Various nomenclature systems are in use, including UniProtKB Accession Numbers, NCBI, Human Genome Organization (HUGO), BioGRID nomenclature, STRING nomen-

	Complete	**Co-expression** (c.s.> 0.4)	**High confidence** (c.s.> 0.7)	**Experimental** (c.s.> 0.4)
Nr. edges	11353056	112776	719552	319892
% edges	100%	0.9%	6.3%	2.8%
Nr. nodes	19576	4499	15131	12984

Tabella 4.1: The table provides essential characteristics of different Human Protein-Protein Interaction (PPI) networks derived from applying various cutoffs to the list of links between human proteins in the STRING database. The networks are classified as follows: "Complete" (no cutoff), "Co-expression" (links related to co-expression with confidence score > 0.4), "High-confidence" (links with confidence score > 0.7), and "Experimental" (links from experimental studies with confidence score > 0.4). The % edges row return the percentage of the number of edges in a PPI network compared to the "Complete" case.

clature, and Gene Ensembl, among others. Tools are available to convert protein names from one nomenclature system to another. However, being able to completely map the all protein names from one nomenclature to another is a very challenging task, far exceeding the purpose of this work.

In the case of the databases extracted from STRING, an attempt was made to convert the protein names from the STRING nomenclature to the NCBI nomenclature. However, the results were not satisfactory, causing a sensible reduction in the dataset dimension. As a result, the decision was made to preserve the STRING nomenclature for all network analyses and perform the conversion only for the discussion section, using the most commonly used names in literature. This approach allowed for accurate referencing and facilitated more comprehensive research when correspondence between naming systems was challenging to establish.

After constructing the human protein-protein interaction (PPI) network, the next step is to extract the virus-host interactions from the original dataset. These interactions refer to edges where either the source or target protein belongs to a virus. Using this procedure, a total of 103 viruses were identified, which is higher than the number present in the BioSTRING dataset. This increase can be attributed to the fact that the STRING database has been updated with additional information since the creation of BioSTRING.

Similar to the previous step, it is necessary to filter the interactions based on a reasonable confidence level. In this case, both high-confidence links related to text mining and medium-confidence links obtained from experimental detection methodologies are considered. By including these additional links, the number of human proteins directly targeted by each virus increases compared to the proteins used in the BioSTRING dataset. This broader inclusion of interactions provides a more comprehensive view of the virus-host relationships in the dataset.

Chapter 5

Bias Detection on Single Layer Networks

5.1 Analysis of Individual Virus-Host Interaction PPI Networks

Before applying the methods of multilayer networks, it is important to analyze each individual virus-host interaction PPI network. This analysis aims to identify the main features, potential differences among viruses, biases that may impact further analysis, and relevant biological details that can guide subsequent investigations.

To verify these observations, it is necessary to create null models. These synthetic PPI networks mimic the characteristics of the dataset but are constructed by randomly selecting nodes according to specific criteria, as outlined in the following section. By comparing the real PPI networks with the null models, we can gain insights into the significance and uniqueness of the observed network properties.

5.1.1 Synthetic Networks

The procedure described in Chapter 3.2.3 for constructing the virus-host PPI network allows us to create nearest neighbors PPI networks based on an input list of nodes, representing the proteins directly targeted by virus proteins.

The choice of the protein list depends on the specific research question at hand. The goal is to extract a certain number of samples from a given protein list, either using a uniform distribution or assigning a probability to each entry. This procedure relies on three main components: the number of samples, the sample list, and the probability distribution.

In this work, the number of samples was determined using one of the following approaches:

- **Fix**: The first approach involves selecting a fixed number of samples, which is suitable when studying the characteristics of a specific virus. Since we cannot control the number of proteins in the virus or the number of nearest neighbors in the human proteome, this fixed number approach ensures consistency across viruses.
- **Distr**: The second approach involves drawing the number of samples from a distribution that matches the empirical distribution observed in the data. As shown in Fig.5.13, the empirical distribution aligns well with an power law distribution, which parameters were find with a fit. Therefore, the number of samples is randomly chosen as an integer extracted from an exponential distribution, which parameters are obtained from a fit of the empirical distribution.

The sample lists used in the following analysis are:

- **Human**: all the proteins in the human PPI.
- **Onco**: all the proteins of the human PPI that are targeted at least by one oncogenic virus.

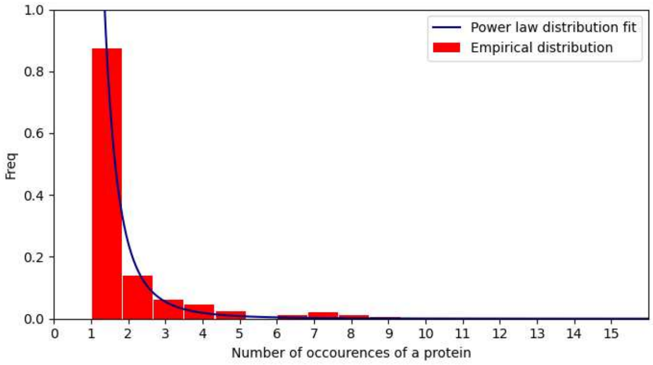


Figura 5.13: The red histogram represents the empirical distribution of the number of human proteins directly targeted by the viruses in the dataset. The blue line represented the power law function fitted on the empirical distribution from which the values to use in the "Distr" nr. of samples determination procedure, are exctracted.

- **Nonco**: all the proteins of the human PPI that are targeted at least by one non-oncogenic virus.
- **Sars**: all the proteins of the human PPI that are targeted by the Sars-Cov2 virus.

The choice of the protein list from which samples are extracted in this work includes the following options:

1. **Unif**: the proteins to use in the synthetic virus are extracted form the sample list with a uniform distribution.
2. **Targeted**: this method aims to create viruses that resemble the original ones more closely. For this reason, it may be appropriate to assign a higher probability to extracting proteins that appear in multiple viruses. To achieve this, a normalized histogram of protein appearances in different viruses can be generated, and each sample can be assigned a probability based on its frequency of occurrence. This method can be used only when the sample set is associated to a empirical distribution associated to the occourcence in different viruses, such as *Onco* and *Nonco*.
3. **Enrich**: This approach involves performing a functional enrichment analysis, typically using KEGG, on the virus of interest. For each detected pathway, the overlap value and the proteins belonging to the pathway are extracted. From each list of proteins, a number of samples are randomly selected using a binomial distribution with the number of elements equal to the number of proteins in the pathway and a probability equal to the overlap value. This selection is performed uniformly across the proteins in the pathway. The aim of this approach is to create viruses that target proteins with similar functional features compared to the original virus. This method has to be used starts from a sample set associated to a specific virus, but at the end the list of samples actual used to build the synthetic virus is different.

For further analysis, the results obtained from the synthetic viruses generated using different methods are compared to identify any significant features that are not mere chance occurrences. Each approach used to create a set is described through the abbreviation:

$$SamplesSet\ \text{-}\ SampelsExctracitionMethod\ \text{-}\ Nr.ofSamplesExctractionMethod$$

5.1.2 Number of Nodes Involved

In the initial analysis, we examine the number of proteins involved in constructing the PPI networks. These proteins can be categorized as follows: virus proteins, human proteins directly targeted by the virus, and human proteins that are nearest neighbors to the directly targeted proteins.

For further analysis, the most relevant quantity among these categories is the number of human proteins directly targeted by the virus, as it determines the size of the virus-related PPI network. Such quantities are graphically represented in Fig.5.14. Statistical analysis reveals that oncogenic

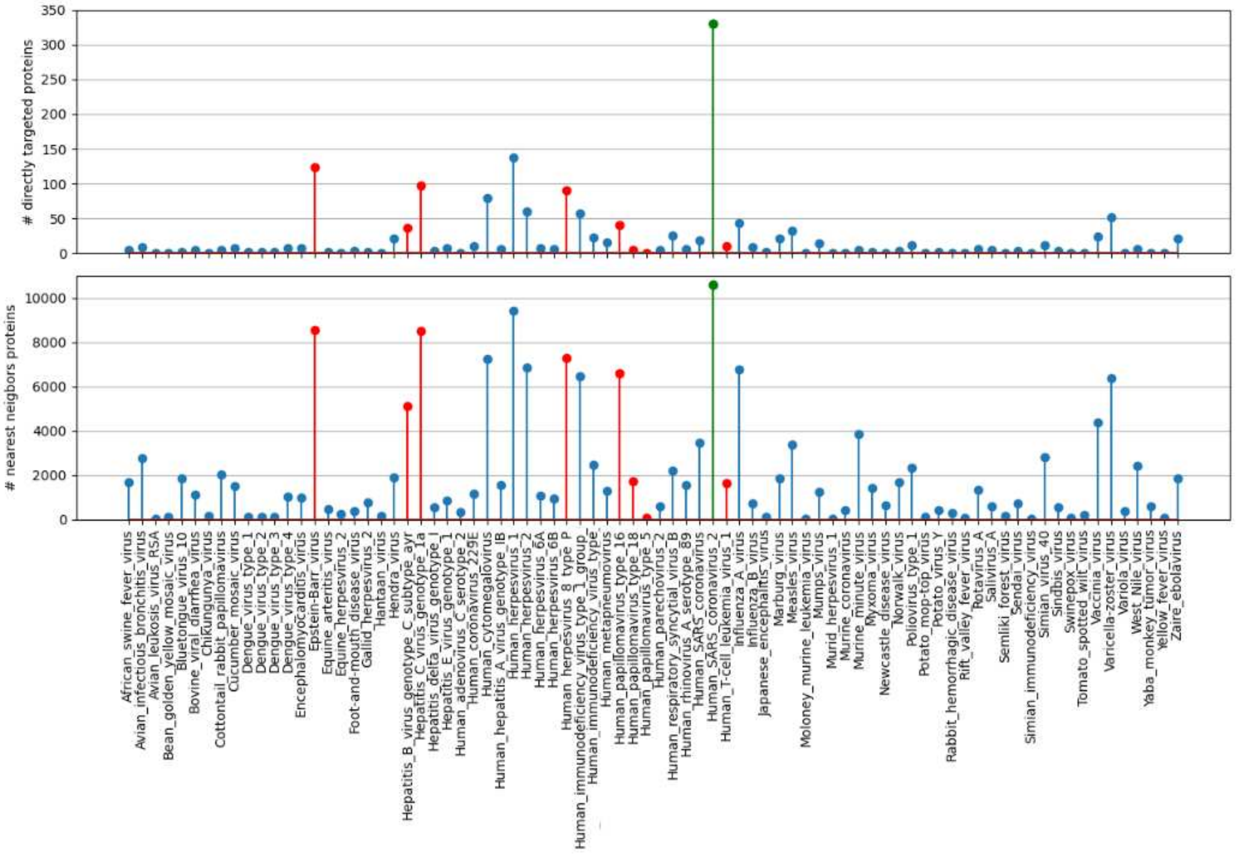


Figura 5.14: The stem plots display two sets of data: the number of human proteins directly targeted by viruses (top) and the number of nearest neighbors of these proteins in the Human PPI network (bottom). Each virus in the dataset is represented, with oncogenic viruses shown in red, non-oncogenic viruses in blue, and the Sars-Cov2 virus highlighted in green.

viruses have a larger number of directly targeted human proteins (50 ± 44) compared to non-oncogenic viruses (12 ± 21). Interestingly, the Sars-Cov2 virus exhibits the highest number of directly targeted human proteins. To determine if this value is solely due to the extensive research conducted on this specific virus or if it genuinely interacts with a large number of human proteins, it is necessary to compare it with synthetic viruses generated using a specific procedure.

The network sizes of 100 synthetic viruses, created using various procedures, are extracted and compared. The procedures include: *Human-Unif-Fix(Sars)*, *AllViruses-Targeted-Fix(Sars)*, *Sars-Enrich-Fix(Sars)*, *Sars-Unif-Distr*, *Sars-Enrich-Distr*, *Nonco-Targeted-Distr*, *Onco-Targeted-Distr*, *Human-Unif-Distr*, and *AllViruses-Targeted-Distr*. The results are presented in Fig.5.15, alongside the network size values of the viruses in the dataset, by also dividing them in the oncogenic and non-oncogenic groups. In order to better compare the different distributions, are also computed the p-values associated to the application of the Wilcoxon-Ranksum test which verifies the hypothesis of two distributions having the same median.

Upon analyzing the results, several observations can be made:

- The distribution *Sars-Enrich-Fix* is the only one that is compatible with the size of the Sars-Cov2 PPI network. This is reasonable because the targeted proteins are extracted based on the pathways associated with the Sars-Cov2. Furthermore, since the proteins to target in the synthetic viruses are chosen uniformly from the list of proteins associated with each pathway, it can be concluded that the Sars-Cov2 does not exclusively target the most important and highly connected proteins within each pathway.

- The values in the *Sars-Unif-Distr* distribution are statistically smaller than the size of the Sars-Cov2 network. This indicates that the virus targets a set of nodes with a higher degree than the

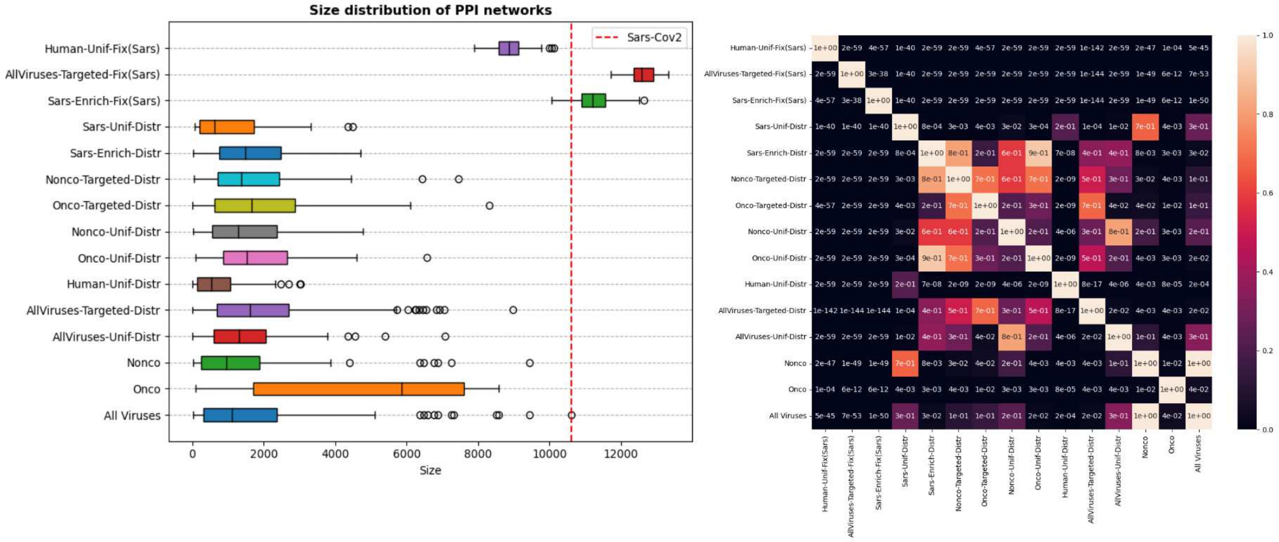


Figura 5.15: On the left, istribution of the networks size for different sets of PPI networks. They involve viruses PPI from the dataset (All Viruses, Onco, Nonco) but also synthetic viruses build with different methods. The red dashed lines represent the Sars-Cov2 PPI network size. On the right, heatmap indicating the p-values obtained by comparing pairs of distributions using the Wilcoxon-Ranksum test.

average value.

- The values in the *AllViruses-Targeted-Fix(Sars)* distribution are statistically higher than the size of the Sars-Cov2 network. This suggests that although other viruses interact with a smaller number of human proteins, these proteins are more connected in the human interactome.
- Comparing *Sars-Unif-Distr*, *Human-Unif-Distr*, and *AllViruses-Targeted-Distr*, it can be observed that *Human-Unif-Distr* and *Sars-Unif-Distr* are compatible with each other, while the values in *AllViruses-Targeted-Distr* are statistically higher. This supports the statement that most of the proteins targeted by the Sars-Cov2 are not as highly connected as the human proteins targeted by other viruses.
- The distributions *Nonco-Unif-Distr*, *Nonco-Targeted-Distr*, *Onco-Unif-Distr*, *Onco-Targeted-Distr*, *AllViruses-Unif-Distr*, and *AllViruses-Targeted-Distr* appear to be compatible with each other. This suggests that proteins targeted a greater number of times are not necessarily more highly connected in the human interactome compared to others. The *Sars-Enrich-Distr* distribution can also be grouped with these distributions, possibly because the Enrich method is more likely to extract pathways that are more relevant in the Sars-Cov2 enrichment analysis. One possible explanation is that these pathways are particularly important for the cell and therefore composed of highly connected proteins.

The initial analysis indicates that Sars-Cov2 exhibits structural properties that distinguish it significantly from other viruses, both oncogenic and non-oncogenic. This divergence may pose challenges when comparing Sars-Cov2 with viruses from the two classes after the classification procedure.

One potential solution is to reduce the dimensionality of the associated PPI network for Sars-Cov2. This can be achieved by increasing the confidence level threshold for retaining specific interactions or by conducting bootstrap analyses. In the case of bootstrap analysis, elements belonging to the most relevant pathways could be extracted more frequently or in a uniform manner. These approaches aim to mitigate the structural differences and facilitate meaningful comparisons between Sars-Cov2 and viruses from other classes.

Chapter 6

Topological Features Analysis on Multilayer Networks

The approach employed in this study involves constructing multilayer networks by stacking multiple protein-protein interaction (PPI) networks corresponding to different viruses. A total of four layers were chosen to allow for various combinations of oncogenic and non-oncogenic viruses while maintaining manageable system sizes and preserving layer-specific information.

Each node in the network represents a specific protein in the human PPI and is shared across all layers. Based on this, the multilayer network can be constructed using two primary approaches. The first approach is an interconnected multiplex, where each replica of a physical node is connected to all other replicas. If a protein is not present in the nearest neighbor network of proteins directly targeted by a virus in a specific layer, the corresponding replica node in that layer will have inter-layer edges only and no intra-layer links.

The second approach involves an edge-colored network, where interlayer interactions are not explicitly defined. This approach is particularly important for conducting versatility analysis using the multilayer PageRank algorithm, where each node has a probability of uniformly transitioning to nodes in networks not directly connected to it. It is also relevant for the layered DCSBM algorithm used to analyze modules and modularity of the system.

The multilayer framework serves multiple goals. Firstly, it enables the interpretation of different layers as different influences on the same human interactome caused by external stimuli, specifically viral infections in this study. Secondly, by combining multiple systems (virus-host PPI networks) with a shared feature, it becomes possible to identify common properties among individual systems and determine the underlying common feature. Lastly, this framework allows for generating a large number of combinations of four-virus sets, enabling the generation of distributions for each quantity extracted from the multiplexes, which can be statistically analyzed.

The central idea driving this approach is to initially identify quantities that can differentiate between oncogenic and non-oncogenic viruses and then examine how the inclusion of SarsCov2 in either class affects the results.

The Wilcoxon-Ranksum test will be utilized for comparing distributions, specifically by computing p-values that describe the similarity between pairs of distributions. A threshold of significance at a p-value of $5E-2$ will be employed to determine the rejection of the null hypothesis, which assumes that the two distributions have the same median.

The chosen combinations for analysis are as follows:

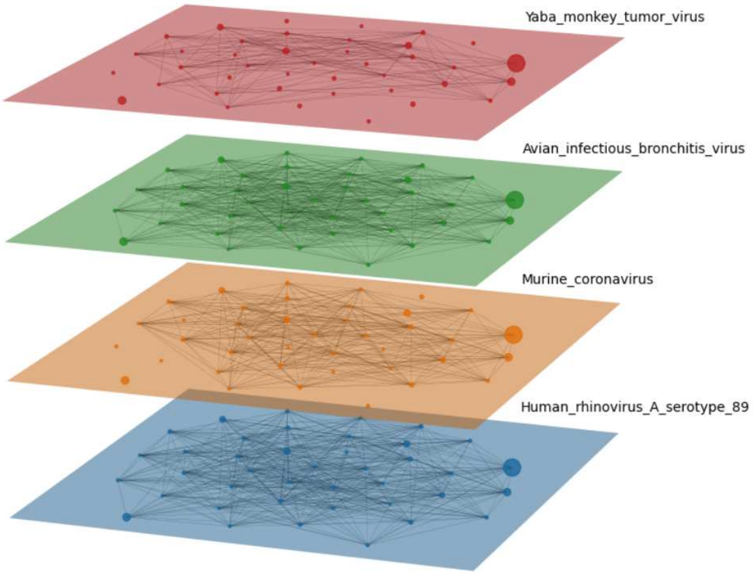

Yaba_monkey_tumor_virus
Avian_infectious_bronchitis_virus
Murine_coronavirus
Human_rhinovirus_A_serotype_89

Figura 6.16: A visual depiction of a multilayer network extracted from a sample in the N combination set. The size of the nodes corresponds to their page-rank centrality within the network.

- Networks composed solely of viruses of one type: oncogenic (O) and non-oncogenic (N). By comparing these two sets, the aim is to identify features that maximize a specific metric between the generated distributions. These features could then be utilized for classification purposes.
- Given the focus on examining the characteristics of individual viruses, it is important to isolate the contribution of each layer in the multilayer network. To achieve this, combinations are constructed starting with viruses of the same category and progressively adding viruses from the other category. These combinations include: 3 non-oncogenic and 1 oncogenic (N1O), 2 non-oncogenic and 2 oncogenic (N2O), and 1 non-oncogenic and 3 oncogenic (N3O).
- Structures are also constructed to specifically study the Sars-Cov2 virus. Sets comprising 3 viruses of the same type along with Sars-Cov2 are created (N1S, O1S). Additionally, other combinations are generated to investigate whether the inclusion of Sars-Cov2 in the network yields a contribution comparable to that of inserting a generic oncogenic virus (N1O1S, N2O1S).

In Tab.6.2 are reported the number of possible combinations for each of the combination sets, alongside to the abbreviation that will be used in the analysis sections.

To compute the required feature distributions for the subsequent analysis, a set of 2000 samples is extracted from all combinations except for O and $O1S$. For these two combinations, all available samples (70 for O and 56 for $O1S$) are included.

6.1 Components Analysis

The initial stage of the multilayer analysis involves examining the largest components of the multiplex network, specifically the largest connected component (LCC), largest intersected component (LIC), and largest viable component (LVC). These quantities were calculated using the functions provided by the `MuxVizPy` package.

The size of these components is the primary focus of this analysis.

6.1.1 LCC

The analysis of the Largest Connected Components (LCCs) confirms that oncogenic PPI networks tend to be larger than non-oncogenic networks. This trend is evident when comparing the distributions of

Set Composition	# of possible combinations	Acronym
4 oncogenic	70	O
4 non-oncogenic	1028790	N
3 non-oncogenic + 1 oncogenic	477120	N1O
2 non-oncogenic + 2 oncogenic	71568	N2O
1 non-oncogenic + 3 oncogenic	4032	N3O
3 oncogenic + SarsCov2	56	O1S
3 non-oncogenic + SarsCov2	59640	N1S
2 non-oncogenic + 1 oncogenic + SarsCov2	20448	N1O1S
1 non-oncogenic + 2 oncogenic + SarsCov2	2016	N2O1S

Tabella 6.2: Description of how the different combination sets are build and the number of possible different samples that is possible to extract from each of them.

oncogenic (O) and non-oncogenic (N) networks in Fig.6.17(left). As more oncogenic layers are added to the multilayer networks, the distributions shift from N to O.

Furthermore, it is evident that all the distributions are highly distinct from each other, as reinforced by observing the heatmap in Fig.6.17. The heatmap clearly shows that only the distributions of $N2O1S$ and $N1O1S$ seem to exhibit compatibility with each other, while the remaining combinations demonstrate significant differences.

The imbalance in network size associated with Sars-Cov2 is also apparent in the three highest distributions ($N1S$, $N2O1S$ and $N1O1S$) and the bottom one ($O1S$) in Fig.6.17, which have values significantly greater than those of O.

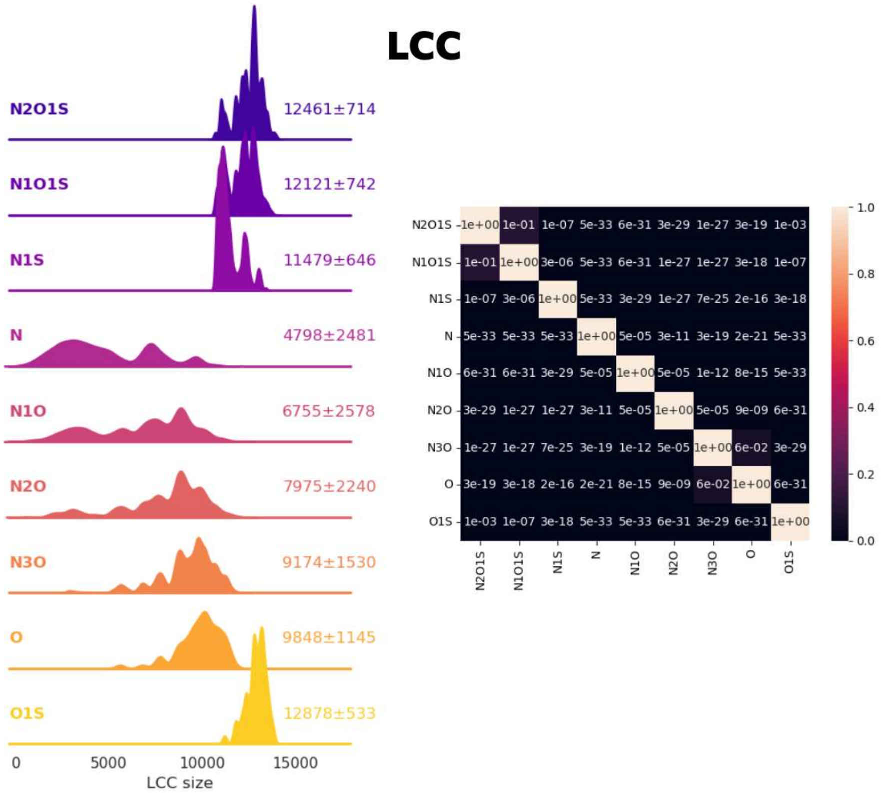
In conclusion, it can be observed that oncogenic viruses generally generate larger PPI networks compared to non-oncogenic viruses, and this effect appears to be even more pronounced for the Sars-Cov2 PPI network.

There are two possible explanations for this observation. First, it is possible that oncogenic viruses have a greater level of interaction with the human PPI network, resulting in larger virus-host PPI networks. Alternatively, this size disparity could be attributed to some form of bias. To address this hypothesis, the subsequent parts of the study will consider this factor and not solely rely on network size for the oncogenic/non-oncogenic classification. By doing so, other potential causes related to various aspects of the network can be explored.

Regarding the composition of the LCC, this analysis is not particularly informative. Since most virus-host PPI networks consist of only one component, and the multilayer framework likely creates connections even between the few disconnected components in each layer, the final result is a union of all the nodes from the four layers.

6.1.2 LIC and LVC

The analysis of the Largest Intersected Components (LICs) allows us to investigate common features among viruses within the same multilayer network, particularly by identifying a shared core of nodes involved in the PPIs of viruses in the same category. Additionally, the LIC can be compared with the corresponding Largest Viable Component (LVC), which is a subset of the LIC and represents the set of nodes that are simultaneously connected by the same path in all layers. The comparisons between LIC and LVC sizes for different combination sets are shown in Fig.6.18, alongside the heatmaps reporting the p-values associated to the compatibility Wilcoxon-Ranksum tests between the different distributions. In the graphical representation, the size of LIC and LVC are treated as a single quantity. This is because there is no visible distinction between the two when examining the graphs. Similarly, the statistical descriptive measures such as median and standard deviation also show no noticeable difference between LIC and LVC sizes.



	N2O1S	N1O1S	N1S	N	N1O	N2O	N3O	O	O1S
N2O1S	1e+00	1e-01	1e-07	5e-33	6e-31	3e-29	1e-27	3e-19	1e-03
N1O1S	1e-01	1e+00	3e-06	5e-33	6e-31	1e-27	1e-27	3e-18	1e-07
N1S	1e-07	3e-06	1e+00	5e-33	3e-29	1e-27	7e-25	2e-16	3e-18
N	5e-33	5e-33	5e-33	1e+00	5e-05	3e-11	3e-19	2e-21	5e-33
N1O	6e-31	6e-31	3e-29	5e-05	1e+00	5e-05	1e-12	8e-15	5e-33
N2O	3e-29	1e-27	1e-27	3e-11	5e-05	1e+00	5e-05	9e-09	6e-31
N3O	1e-27	1e-27	7e-25	3e-19	1e-12	5e-05	1e+00	6e-02	3e-29
O	3e-19	3e-18	2e-16	2e-21	8e-15	9e-09	6e-02	1e+00	6e-31
O1S	1e-03	1e-07	3e-18	5e-33	5e-33	6e-31	3e-29	6e-31	1e+00

Figura 6.17: On the left the normalized distribution plots depict the LCC size for samples belonging to different combination sets. The median and standard deviation values are provided for each distribution, offering insights into the central tendency and variability of the respective variables. On the left the heatmap displays the p-values obtained from the Wilcoxon-Ranksum test applied to all possible pairs of distributions.

Interestingly, the size distributions of oncogenic viruses (O) are consistently larger than those of non-oncogenic viruses (N) in both the LIC and LVC cases. This suggests that there might be a specific set of nodes within the PPI network that predominantly interact with oncogenic viruses, leading to their larger size. On the other hand, non-oncogenic viruses seem to have a more diffuse and less concentrated influence on a broader range of proteins, resulting in smaller and less distinct size distributions.

Consequently, it is highly informative to examine the specific proteins that constitute both the Largest Intersected Components (LICs) and Largest Viable Components (LVCs). By intersecting all the LICs and LVCs within a particular combination set, we can identify the common core of nodes shared among the viruses in that set. This analysis is particularly interesting for both the oncogenic (O) and non-oncogenic (N) cases. When applying the intersection procedure to O, we find that the intersection for both the LIC and LVC yields the same set of nodes:

RNF4, DAXX, PARP1, TP53, MDM2, CREBBP, ABL1, HMGA1, HIST2H2BE,
PIN1, FBXW7, PML, TRIM25, MAP3K1, MDM4, UBE2I, NKX2-1, SMAD3, TP73,
PPM1D, HNRNPL, KIAA1429, SMAD2, CTBP1, CBX4, ACTBL2, RANBP9,
SUMO2, SKI, PIAS1

On the other hand, for the non-oncogenic viruses (N), no common core is observed when intersecting all the LICs and LVCs. This lack of common core could be attributed to the larger number of non-oncogenic viruses compared to oncogenic ones, reducing the likelihood of a shared set of connected nodes. Additionally, it is worth noting that the PPI networks associated with non-oncogenic viruses are statistically smaller than those of oncogenic viruses (Fig.5.14).

However, it is still informative to investigate which proteins appear more frequently in the LIC and LVC node sets for N. The results are presented in Tab.6.3, highlighting the difference in occurrences between the LIC and LVC for the most frequent proteins. Notably, the proteins *HSP90AA1* and

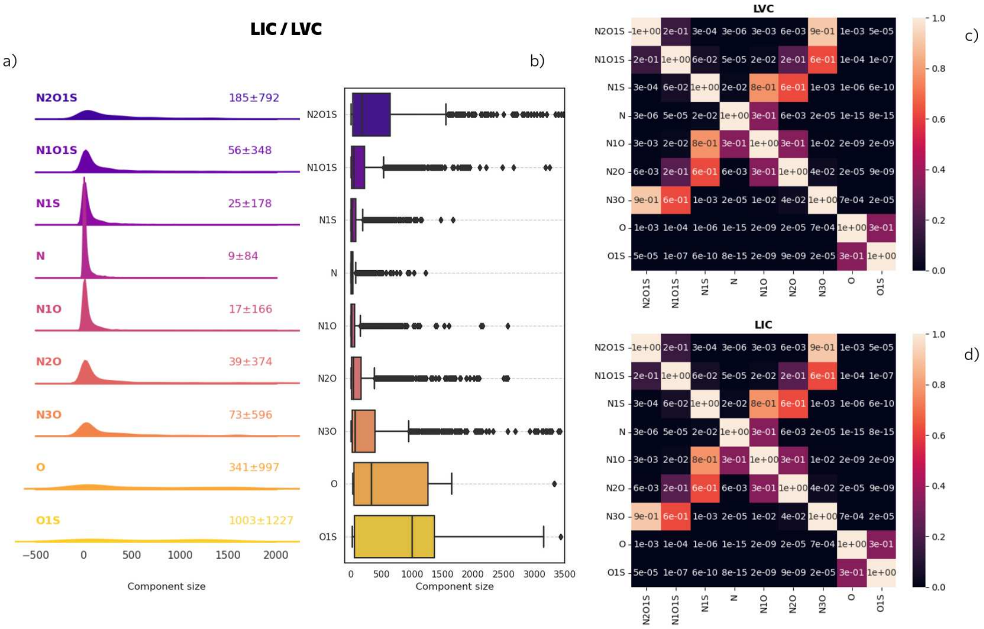


Figura 6.18: Figure (a) shows boxplots and figure (b) the normalized distribution plots of the LIC and LVC size distributions for samples from different combination sets. The median and standard deviation values are provided for each distribution in (b). It can be observed that there is no significant difference in the distribution of LIC and LVC sizes, both graphically and statistically, as indicated by the lack of noticeable differences in descriptive quantities. This observation is further supported by the heatmaps in (c) and (d), which display the p-values obtained from the Wilcoxon-Ranksum test applied to all possible pairs of distributions in the two cases.

CDK9 exhibit the greatest percentage difference in their frequencies. These proteins are involved in fundamental cellular processes, suggesting that in the layers where they are not directly connected, the global connectivity tends to establish alternative paths to connect them to the largest component.

Upon examining the heatmaps in Fig.6.18, an interesting observation can be made regarding the compatibility between certain distribution pairs. Notably, the following pairs exhibit compatibility: O and $O1S$, $N1O$ and $N1S$, $N2O$ and $N1O1S$, $N3O$ and $N2O1S$. This indicates that when substituting a layer associated with an oncogenic virus with the one associated with Sars-Cov2, the resulting distributions of LIC and LVC size remain compatible with the original distributions. This suggests a strong affinity between Sars-Cov2 and oncogenic viruses. Additionally, compatibility is observed between $N3O$ and $N1O1S$, as well as between $N2O$ and $N1S$, indicating that the inclusion of the Sars-Cov2 layer seems to increase the "degree of oncogenicity" even more compared to the introduction of an oncogenic layer.

6.2 Percolation Analysis

Another approach to analyze the properties of the systems is to examine their robustness to percolation processes. Percolation, specifically nodes percolation, involves progressively removing nodes from the networks. As described in the previous chapter, the order in which nodes are removed, along with their corresponding links, in the multilayer networks is determined by their descending order of pagerank versatility when considering the system as an edge-colored network.

In this analysis, the focus is on the critical point, which refers to the fraction of removed nodes at which a peak in the dimension of the second largest component is observed. This critical point provides insights into the network's resistance to targeted attacks, which, from a biological perspective, could correspond to the action of certain drugs.

The set of combinations used for this analysis is the same as mentioned in the Components section. The critical points generate different distributions, as shown in Fig. 6.19. By examining the image, it

Protein	Abs Freq LVC	Abs Freq LIC	Abs Freq Diff	% Freq Diff
TP53	796	799	3	0.37
RELA	733	734	1	0.13
MYC	643	649	6	0.92
TNF	643	646	3	0.46
NTRK1	601	612	11	1.79
TRIM25	570	570	0	0
UBC	502	502	0	0
COPS5	463	464	1	0.21
HSP90AA1	406	443	37	8.35
CTNNB1	422	422	0	0
NFKB1	409	409	0	0
RPS27A	400	400	0	0
AKT1	395	395	0	0
KIAA1429	365	365	0	0
CUL1	356	362	6	1.65
APP	354	356	2	0.56
KRAS	338	339	1	0.29
UBA52	329	329	0	0
ESR1	327	328	1	0.30
JUN	317	317	0	0
CDK9	300	313	13	4.15

Tabella 6.3: The table presents the occurrence counts in the LICs and LVCs of 2000 samples from the N combination set, of the 20 proteins which appears more frequently in the LICs of the considered samples. It aims to highlight proteins that exhibit a substantial percentage difference in occurrence between LICs and LVCs.

is apparent that when transitioning from N to O by adding oncogenic layers, the distribution changes from being widely spread, particularly with a thick and long tail for low values of the critical point (ranging from 0.55 to 0.65), to progressively reducing the dimension of this tail. The distribution becomes more concentrated around higher critical point values, ultimately resulting in the O distribution characterized by a significantly smaller variance.

Based on the observations, it can be concluded that multilayer networks composed of oncogenic viruses exhibit greater statistical resistance to such attacks compared to systems belonging to N. Furthermore, the responses of these oncogenic networks are characterized by less fluctuation in intensity compared to the non-oncogenic case.

However, the introduction of the Sars-Cov2 virus exhibits a different trend. In particular, when analyzing $N1S$, there is a shift in the peak towards higher values of the critical point, accompanied by a thick tail towards low values. When oncogenic viruses are added, as in $N1O1S$ and $N2O1S$, a double-peak behavior becomes apparent.

Overall, the impact of introducing the Sars-Cov2 virus appears to be distinct from solely introducing oncogenic viruses in multilayer networks with non-oncogenic layers. This discrepancy may be attributed to structural differences between the Sars-Cov2 virus and both oncogenic and non-oncogenic viruses, which could be an interesting area for further investigation.

By examining the compatibility values obtained from the Wilcoxon Ranksum test, as depicted in the heatmaps of Fig.6.19, it is evident that the $O1S$ distribution is compatible with $N2O1S$, $N2O$, $N3O$, and O, all of which involve a high number of layers associated with oncogenic viruses.

Of particular interest is the compatibility between $N1S$ and O distributions, which may be attributed to both distributions exhibiting peaks around high values of the critical point. However, this similarity decreases as more oncogenic viruses are introduced into the samples, as observed in the case of $N1O1S$

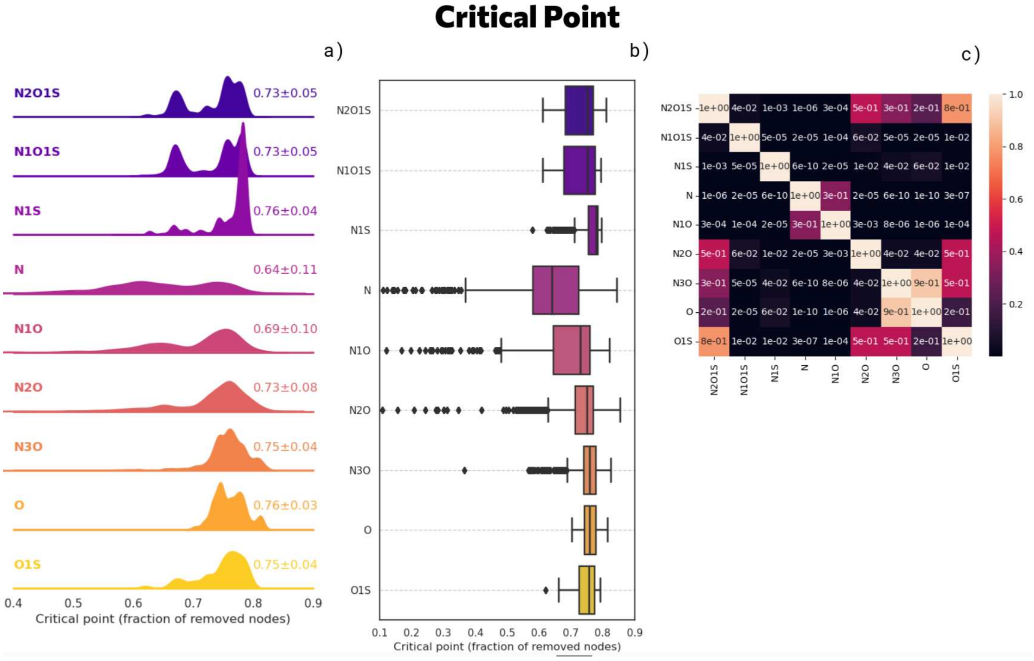


Figura 6.19: Figure (a) shows boxplots and figure (b) the normalized distribution plots of the critical point distributions for samples from different combination sets. The median and standard deviation values are provided for each distribution in (b). The heatmaps in (c) and (d) display the p-values obtained from the Wilcoxon-Ranksum test applied to all possible pairs of distributions.

and $N2O1S$, where a double-peak behavior reminiscent of the N distribution reemerges, however in this last the effect is less evident.

6.3 Modules and Modularity

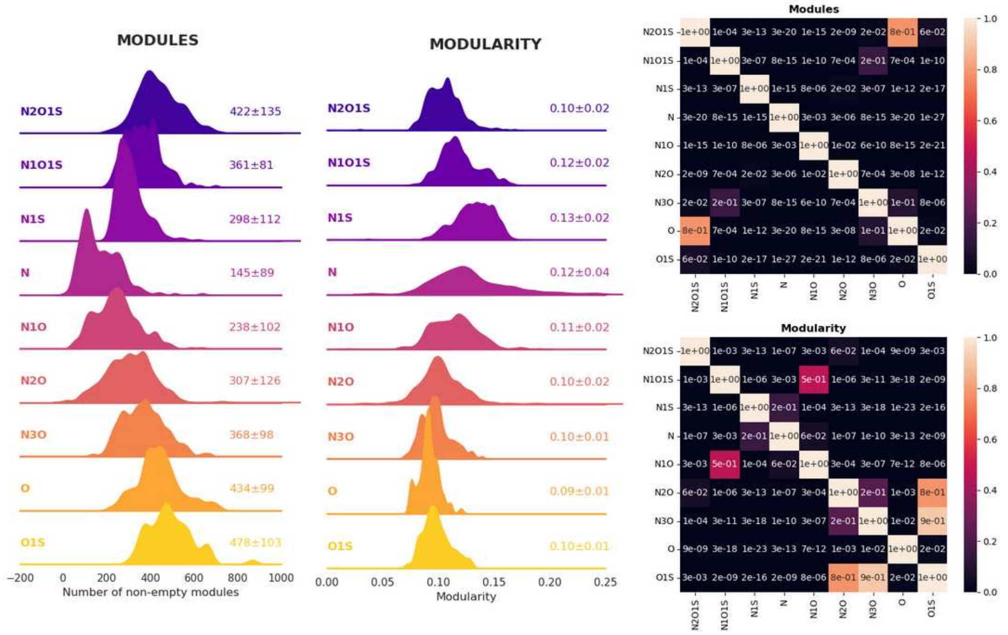
The mesoscale structure of the multilayer networks is further explored by comparing the results of the degree corrected stochastic block model. The 2 analyzed quantities are: the number of non-empty modules and the modularity of the systems.

The results are presented in Fig.6.20, alongside to the heatmaps with the p-values associated to the Wilcoxon-Ranksum tests used to compare the different distributions.

The addition of oncogenic viruses produces a statistical significant increase of the number of non-empty modules, and this effect is further amplified when the Sars-Cov2 virus is included. Conversely, the modularity values tend to decrease when oncogenic viruses are added to non-oncogenic combinations. However, before drawing conclusions based on these quantities, it is important to examine their correlation with network size. Previous sections have shown that oncogenic layers generally have larger PPI networks compared to non-oncogenic ones.

To investigate this correlation, scatterplots were created, plotting the size of the largest connected component (LCC), which is related to network size, against the number of non-empty modules and the modularity. In Fig.6.21, it can be observed that there is a moderate positive correlation between network size and the number of non-empty modules (linear correlation coefficient = 0.65). Thus, the number of non-empty modules may not provide substantial information, in fact high values of such a quantity are will be sign of larger networks rather than informations about specific pattern in the distribution of communities.

In contrast, the relationship between modularity and LCC size does not demonstrate a strong linear pattern, as evidenced by the low linear regression coefficient (-0.18). Moreover, the observed trend appears to be opposite to that of the number of modules. These findings suggest that modularity, as



Modules

	N2O1S	N1O1S	N1S	N	N1O	N2O	N3O	O	O1S
N2O1S	1e+00	1e-04	3e-13	3e-20	1e-15	2e-09	2e-02	8e-01	6e-02
N1O1S	1e-04	1e+00	3e-07	8e-15	1e-10	7e-04	2e-01	7e-04	1e-10
N1S	3e-13	3e-07	1e+00	1e-15	8e-06	2e-02	3e-07	1e-12	2e-17
N	3e-20	8e-15	1e-15	1e+00	3e-03	3e-06	8e-15	3e-20	1e-27
N1O	1e-15	1e-10	8e-06	3e-03	1e+00	1e-02	6e-10	8e-15	2e-21
N2O	2e-09	7e-04	2e-02	3e-06	1e-02	1e+00	7e-04	3e-08	1e-12
N3O	2e-02	2e-01	3e-07	8e-15	6e-10	7e-04	1e+00	1e-01	8e-06
O	8e-01	7e-04	1e-12	3e-20	8e-15	3e-08	1e-01	1e+00	2e-02
O1S	6e-02	1e-10	2e-17	1e-27	2e-21	1e-12	8e-06	2e-02	1e+00

Modularity

	N2O1S	N1O1S	N1S	N	N1O	N2O	N3O	O	O1S
N2O1S	1e+00	1e-03	3e-13	1e-07	3e-03	6e-02	1e-04	9e-09	3e-03
N1O1S	1e-03	1e+00	1e-06	3e-03	5e-01	1e-06	3e-11	3e-18	2e-09
N1S	3e-13	1e-06	1e+00	2e-01	1e-04	3e-13	3e-18	1e-23	2e-16
N	1e-07	3e-03	2e-01	1e+00	6e-02	1e-07	1e-10	3e-13	2e-09
N1O	3e-03	5e-01	1e-04	6e-02	1e+00	3e-04	3e-07	7e-12	8e-06
N2O	6e-02	1e-06	3e-13	1e-07	3e-04	1e+00	2e-01	1e-03	8e-01
N3O	1e-04	3e-11	3e-18	1e-10	3e-07	2e-01	1e+00	1e-02	9e-01
O	9e-09	3e-18	1e-23	3e-13	7e-12	1e-03	1e-02	1e+00	2e-02
O1S	3e-03	2e-09	2e-16	2e-09	8e-06	8e-01	9e-01	2e-02	1e+00

Figura 6.20: The two normalized distribution plots on the right depict the number of non-empty modules and modularity for samples belonging to different combination sets. The median and standard deviation values are provided for each distribution, offering insights into the central tendency and variability of the respective variables. On the left side are reported the heatmaps with the p-values associated to the Wilcoxon-Ranksum compatibility tests for each possible pair of distributions for number on non-zero modules (top) and modularity (bottom).

a feature of the multilayer networks, is not heavily influenced by network size and may serve as an informative characteristic for distinguishing between oncogenic and non-oncogenic viruses.

Upon examining the figures, it is apparent that the distribution associated with oncogenic viruses (O) is statistically smaller than that of non-oncogenic viruses (N). This discrepancy may indicate that oncogenic viruses exhibit a more dispersed network structure, potentially due to their broader targeting of the human PPI network, as opposed to the more focused approach of non-oncogenic viruses.

To further analyze the differences between the distribution of modularity and the number of non-empty modules for different combination sets, the distances between the distributions are computed using the Wilcoxon rank-sum test. This non-parametric statistical test is suitable for comparing distributions of independent samples when the assumptions of normality or equal variances required by parametric tests, such as the t-test, are not met. Th null hypothesis consists in supposing that the medians of the two distributions under analysis coincide. The matrices reporting the p-values associated to such tests are reported in Fig.6.20.

Upon examining the modularity distributions, an intriguing pattern emerges. Interestingly, certain pairs of distributions exhibit compatibility: $N1S$ and N, $N1O1S$ and $N1O$, $N2O1S$ and $N2O$, $O1S$ and $N3O$. When a non-oncogenic virus layer is substituted with the PPI network associated with Sars-Cov2, the resulting distributions become compatible with each other. This analysis suggests that Sars-Cov2 shares some similarities with non-oncogenic viruses.

In contrast, the number of non-zero modules distributions display substantial differences, which is further confirmed by the heatmap of p-values. Notably, the distributions O and $N2O1S$ are compatible, as well as $N3O$ and $N1O1S$. Considering the correlation with network size, it can be concluded that the shift towards higher values in the number of modules resulting from the introduction of the Sars-Cov2 layer is comparable to the effect of introducing two oncogenic virus layers. This supports the hypothesis that the number of non-zero modules may not be as informative as modularity in this context.

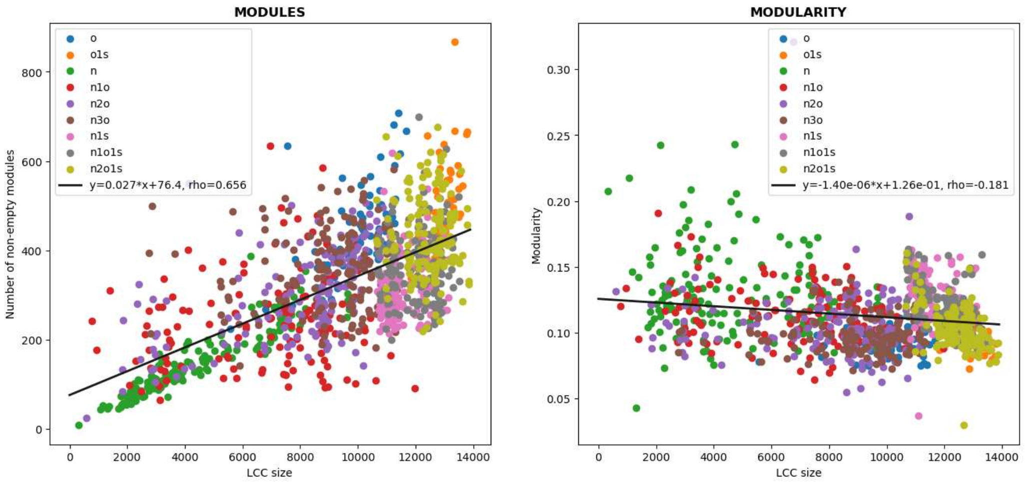


Figura 6.21: The scatterplot illustrates the relationship between the LCC size of sets of multilayer network samples from different combination sizes and two quantities: the number of non-empty modules (right) and modularity (left). Each plot includes a regression line (in black), and the linear correlation coefficient is provided in the legend. This analysis aims to explore any potential correlation between the LCC size (indicative of network size) and the two variables of interest.

6.4 Features Separation

After obtaining information about the topological features that characterize multilayer networks belonging to different combination classes, the next step is to explore the possibility of performing classification based on a specific set of extracted topological features from a multilayer network. These features include LCC (Largest Connected Component), LIC (Largest Independent Component), LVC (Largest Vertex Component), number of non-empty modules, modularity, and the critical point from the node percolation. The dataset consists of 6-dimensional samples.

To visualize and perform the classification task, the multidimensional space where each sample resides is projected onto a 2D plane using UMAP, a dimensionality reduction technique that aims to preserve the global structure of high-dimensional data in a lower-dimensional space. Figure 6.22 shows the result of UMAP dimensionality reduction using all the aforementioned features as input. To perform the classification, a Support Vector Machine (SVM) with Gaussian Kernels is utilized to separate the 2-dimensional UMAP space into two regions associated with the O and N samples. The SVM is trained only with samples belonging to these two combination sets, labeled as 0 and 1, respectively.

The top-left block of the figure displays the N and O sample points used for classification, along with the SVM model generated for this task. The other blocks represent the UMAP representation of the remaining combination sets described earlier. In the top-left corner of each block, the fraction of samples classified as belonging to the O class is reported.

Observing the images related to $N1O$, $N2O$, and $N3O$, it can be noticed that the portion of points classified as belonging to the O class increases as the number of oncogenic layers in the network increases.

On the other hand, when examining samples that include Sars-Cov-2 layers, namely $N1S$ and $O1S$, it is evident that 100% of their samples are classified as O. However, it is also apparent that the region where these points are located is relatively isolated from the other points. This could be attributed to the fact that some of the input features, particularly the dimension of the Largest Connected Component, exhibit a distinct behavior that highly depends on the network's dimension.

To mitigate this effect, we excluded such features from the samples. Additionally, considering that the dimension of LIC and LVC are usually the same in the networks we are analyzing, only the LIC dimension is retained. Figure 6.23 presents the results of the same classification procedure but without considering these two input features.

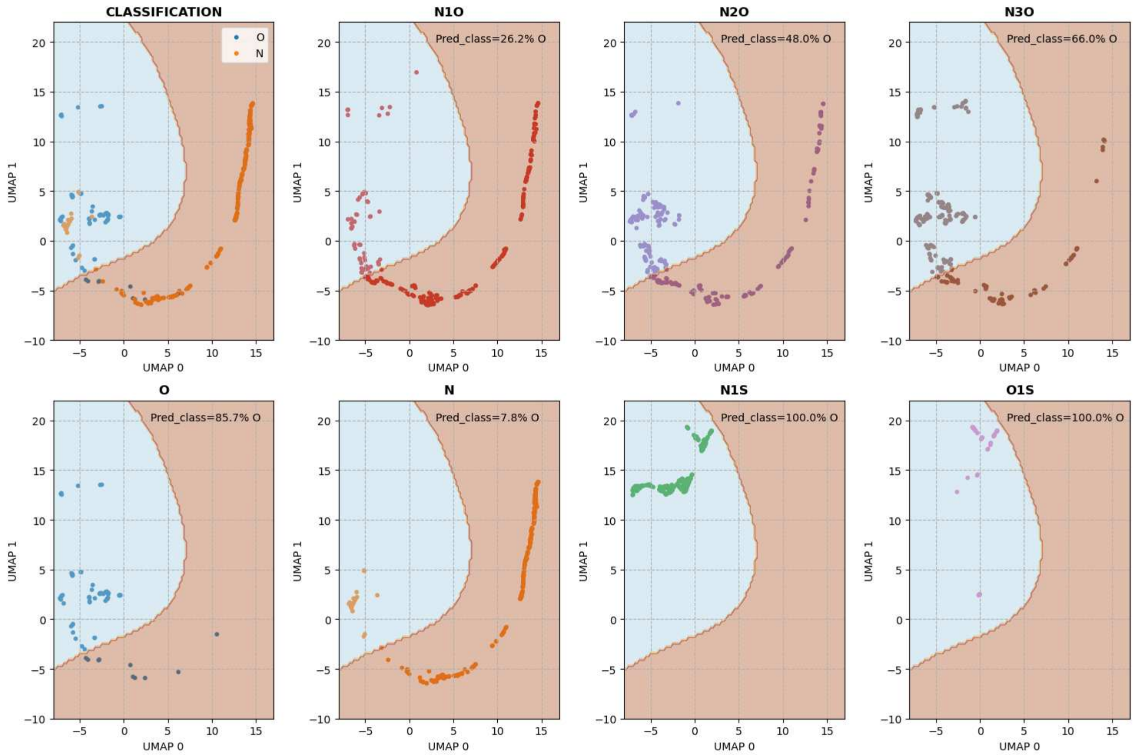


Figura 6.22: The two-dimensional UMAP representation displays the 6-dimensional vector of features (LCC size, LIC size, LVC size, percolation critical point, number of non-empty modules, and modularity) derived from samples belonging to different combination sets. The parameter space is divided into two regions based on the classification performed by the Gaussian kernel SVM between samples belonging to class O (oncogenic) and class N (non-oncogenic). Each image corresponding to a combination set also includes the percentage of samples classified as belonging to the O class, displayed in the top right corner.

The outcome of this classification algorithm clearly demonstrates that the reduced parameter space exhibits distinct separable regions and also that similar results can be reached even with a simpler linear classification approach. Furthermore, by adding more oncogenic layers from the N set, the percentage of samples classified as belonging to the oncogenic region increases, indicating good classification performance.

It is important to note that all samples, including those from $N1S$ and $O1S$, occupy regions in space that are also populated by points from other combination sets, in contrast to the previous. Consequently, it's possible to consider to extend the conclusions drawn from the classification procedure to samples containing the Sars-Cov-2 PPI layer. Specifically, it is evident that the distribution of points from O and $O1S$ sets is very similar, as is the case for $N1O$ and $N1S$ sets. On the other hand, supposing that the Sars-Cov2 can show similar features to the ones associated to the non-oncogenic viruses, it should be observed a similarity between the classification of the N and $N1S$ samples. In the first case the number of samples classified as O are only 6.2%, while in the other case the value raises until 37.9%, results that is hard to explain as a noise effect.

This suggests that the classification algorithm treats Sars-Cov-2 in a comparable manner to the viruses in the oncogenic set.

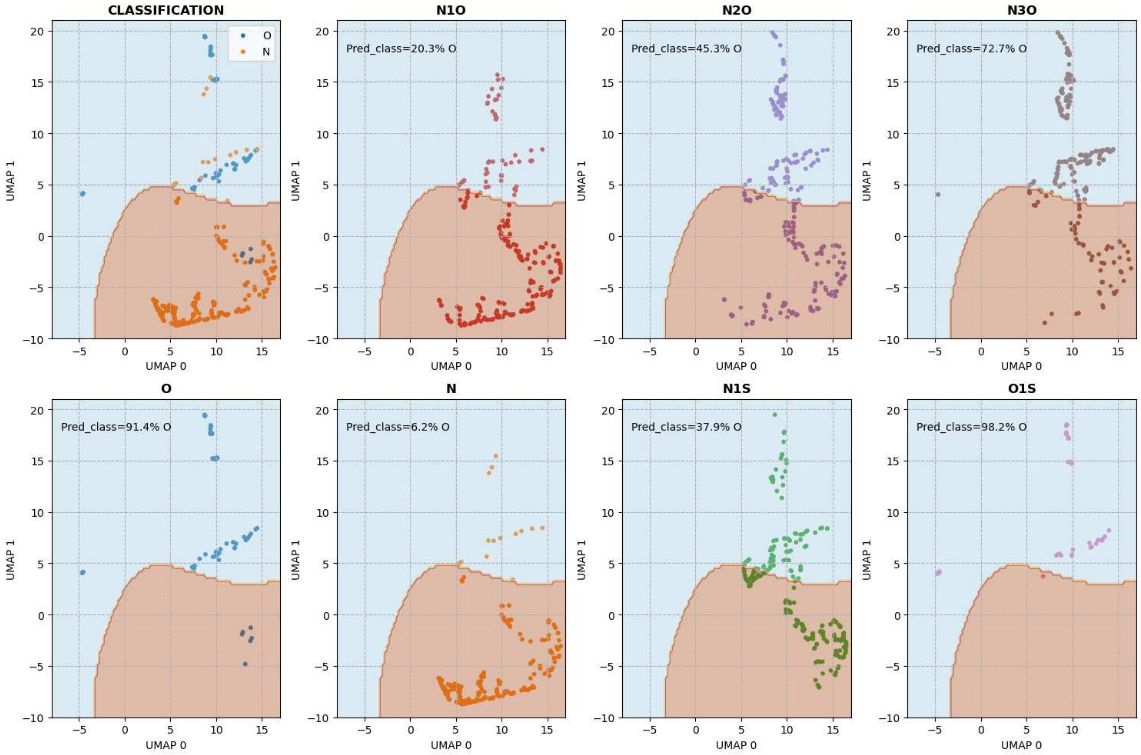


Figura 6.23: The two-dimensional UMAP representation displays the 4-dimensional vector of features (LIC size, percolation critical point, number of non-empty modules, and modularity) derived from samples belonging to different combination sets. The parameter space is divided into two regions based on the classification performed by the Gaussian kernel SVM between samples belonging to class O (oncogenic) and class N (non-oncogenic). Each image corresponding to a combination set also includes the percentage of samples classified as belonging to the O class, displayed in the top right corner.

Chapter 7

Classification with Machine Learning

After analyzing some features of the PPI networks to identify significant differences between oncogenic and non-oncogenic viruses, the next step is to directly address the classification task.

It can be very hard to identify unique characteristics related to the oncogenic nature of a particular virus PPI by inspecting general network features. Tumor development is a complex process, and targeting the human proteome can vary depending on the type of cancer, which is not considered in this work. Moreover, viruses with similar effects on the human body, such as respiratory viruses, intestinal viruses, skin viruses, and hepatitis, can have similar PPIs[43].

To propose a classification method, it is necessary to leverage the prior knowledge we have about the data. This entails using supervised algorithms with interpretable structures to train and analyze them for key features useful in classification[28]. The following sections present the methods and results for two such methods: random forest and perceptron, a neural network without hidden layers.

For both methods, an input and a classification task need to be chosen. Considering the multilayer framework used in this work and the combination sets used for analysis, a possible classification task could be detecting whether one of the layers in the network corresponds to a PPI network of an oncogenic virus. This reduces the problem to binary classification, where the two possible outcomes answer the question, ”Does the network contain an oncogenic layer?” The chosen combination sets are, therefore, N and $N1O$.

We maintain the 4-layer multilayer framework to generate an appropriate number of samples for training the models while adequately concealing the contribution of individual layers during multilayer analysis. This prevents the algorithm from simply identifying the 8 layers we labeled as oncogenic, and instead focuses on finding features that genuinely encode the oncogenic property, which can be generalized to unknown virus PPIs.

After setting up the dataset properly to maximize this characteristic, the model can be trained, enabling the prediction of labels for new samples. At this point, combinations involving the Sars-Cov2 PPI network are considered, and predictions are made to determine its belonging to the oncogenic class or the non-oncogenic class.

Regarding the input of the model, multiple choices are possible, but in this work, two methods are presented.

Firstly, a global representation of the multilayer is aimed for. PageRank versatility is computed for all nodes in an edge-colored multilayer network. An empty vector is created, with each entry mapped to a specific protein in the human proteome. The centrality values obtained earlier are associated with their corresponding protein in the empty vector, while proteins not present in the multilayer network have a value of 0.

The advantage of this method is that it encodes much information about the topology, nature, and dynamics of the networks. However, a disadvantage is that the final input vectors can still be influenced by the bias of network size. Larger networks will have fewer zero entries, and since network size is proportional to the number of oncogenic virus layers in the network. A method to mitigate this effect is proposed later in the work.

The other method involves extracting functional enrichment analysis from an important set of proteins in the multilayer network.

7.1 Perceptron

The Perceptron method is employed, which is a simple neural network architecture connecting input and output through trainable weights. These weights are updated iteratively to minimize the loss function between the network output and the true label of each sample. Since the task at hand is binary classification, the complexity of the Perceptron model is sufficient to achieve good results. The key advantage of using this model is the ease of interpreting the final results. By analyzing the weight distribution after training, it becomes possible to identify the input features that were most influential in the classification task, thereby extracting a list of proteins likely to play a crucial role in distinguishing between oncogenic and non-oncogenic viruses.

To set up the training, validation, and test sets, a dataset of 5600 samples is used, with 2800 samples from N and 2800 from $N1O$. To ensure the neural network's ability to generalize beyond the specific 8 oncogenic viruses in the dataset, an algorithm is implemented. This algorithm extracts samples from the $N1O$ set that include a specific oncogenic virus, creating a filtered dataset comprising the remaining 7 oncogenic viruses. This filtered dataset is used to generate the training and validation sets, while the $N1O$ samples containing a layer corresponding to the target oncogenic viruses are reserved as the test set. Consequently, the model does not receive input samples from this specific virus during training but it is known a priori that the corresponding samples contain oncogenic virus layers. The training and validation sets are split, with 90% of the dataset used for training and 10% for validation.

Following model training, various test datasets are used to evaluate the final performance, each focusing on a different aspect. These test datasets include:

- **OncoTest**: Consists of samples composed of the layer generated by the virus excluded from the training procedure, aiming to assess the model's generalization of the oncogenic feature. The number of items in this test dataset corresponds to the number of samples in $N1O$ containing the specific oncogenic virus.
- **SarsTest**: Comprises samples containing a layer corresponding to the Sars-Cov2 virus, extracted from the $N1S$ combination set. This test dataset includes 100 items.
- **SyntTest**: Contains samples comprising three layers corresponding to non-oncogenic viruses and one synthetic PPI network created by randomly targeting 330 proteins among those directly targeted by the viruses in the dataset.

These test datasets provide insights into different aspects of the model's performance and allow for a comprehensive evaluation.

7.1.1 Model Definition and Training

The model is a neural network architecture with one input layer consisting of $n_{in} = 19945$ neurons and an output layer with $n_{out} = 2$ neurons. The connection between these two layers is of the feed-forward type, resulting in a total of $(n_{in} + 1) \cdot n_{out} = 39892$ trainable parameters.

However, the dataset constructed in this way still faces the issue of sample size. We know that larger multilayers, indicated by samples with a greater number of non-zero entries, are likely associated with

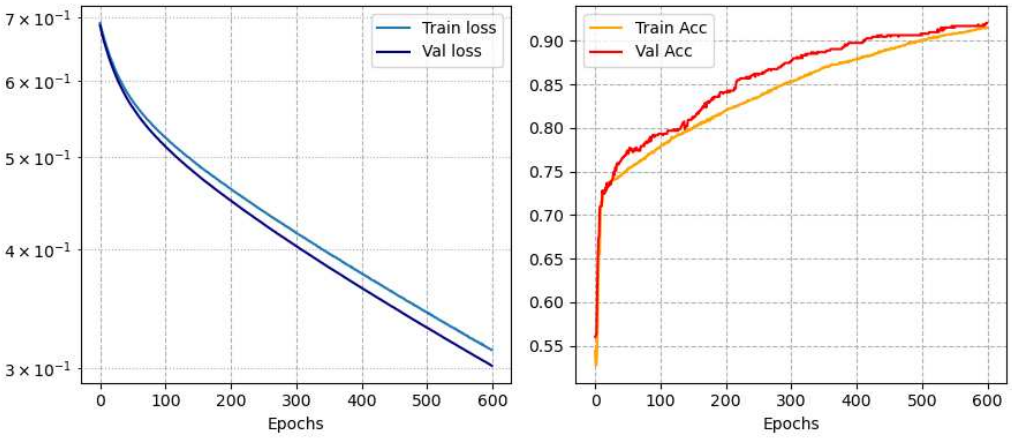


Figura 7.24: Representation of the trend of a Perceptron model's loss (on the right), and accuracy (on the left), computed on the trianing and validation dataset

the presence of layers corresponding to oncogenic viruses. Consequently, the model might learn to separate large networks from smaller networks, which is not the intended purpose of this work.

To address this problem, a sample processing step is proposed. Only the largest 2000 non-zero entries of each sample are retained, while the remaining entries are set to 0. The final vector is then normalized, ensuring values range from 0 to 1. For samples with a number of non-zero entries smaller than 2000, $2000 - n$ random entries are assigned random values. This ensures that the sum of the original n non-zero entries corresponds to 85% of the sum of all entries, while the sum of the random $2000 - n$ entries corresponds to the remaining 15%.

The number 2000 was chosen empirically to retain as much information as possible while preserving key features. Additionally, it ensures that only a small fraction of the input samples have a number of non-zero entries smaller than this threshold (in this case, 8.4

The loss function selected to evaluate and update the model is CategoricalCrossEntropy. Since the model is a classifier between two classes, we want separate weights for each class to aid interpretability. No a priori correction to the class weights was applied to compute the loss, as the number of training samples for each class is comparable. The metric used to evaluate the model is accuracy, which measures the fraction of correctly predicted sample labels.

The Adam optimizer with a learning rate of 5×10^{-4} was chosen.

Regarding the datasets used for training, validation, and testing of the model, several trials were conducted to assess the model's ability to generalize the concept of oncogenic viruses. In each trial, the sample from $N1O$ corresponding to a specific oncogenic virus were excluded from the training procedure and used as a test-set

Each trial involved training the model for 600 epochs, determined empirically as an appropriate number at which the training and validation accuracy reaches a plateau. An example of the training trend can be seen in Figure 7.24. Table 7.4 reports the final training and validation accuracy values, as well as the mean label values calculated from the dataset composed of the $N1O$ samples associated with the test oncogenic virus and the $N1S$ samples. For the first case, all the labels should correspond to 1.

Upon analyzing the table, we observe varying levels of generalization across different trials. The "OncoTest pred" column is particularly informative, as values close to 1 indicate that the model can identify the presence of an oncogenic virus layer in a multilayer network, even if that specific virus was not included in the training. Notably, the best performances were achieved for the removal of "Epstein-Barr virus" and "Human herpesvirus 8 type P" cases. This suggests that training the model with other viruses is sufficient to generalize the problem in these instances.

Trial Name	Excluded Onco Virus	Train Acc	Val Acc	OncoTest pred	SarsTest pred	SyntTest pred
EB	Epstein-Barr	0.923	0.931	0.962	0.911	0.589
HBC	Hepatitis B gen. C, ayr	0.916	0.9204	0.670	0.972	0.617
HC1	Hepatitis C gen. 1a	0.931	0.920	0.886	0.900	0.400
HV8P	Hum. herpesvirus 8 type P	0.923	0.898	0.969	0.933	0.657
PV16	Hum. papillomavirus type 16	0.918	0.911	0.736	0.960	0.620
PV18	Hum. papillomavirus type 18	0.917	0.916	0.833	0.970	0.740
PV5	Hum. papillomavirus type 5	0.932	0.932	0.402	0.953	0.640
TL1	Hum. T-cell leukemia 1	0.915	0.920	0.860	0.978	0.667

Tabella 7.4: The table presents the performance results of individual Perceptron models trained using datasets in which samples containing specific oncogenic viruses were excluded. Each trained model was subsequently tested on different sample sets: OncoTest consists of samples containing the excluded oncogenic virus PPI layer, SarsTest contains samples with the SARS-CoV-2 PPI network, and SyntTest comprises synthetic layers composed of the nearest neighbors of 330 proteins randomly selected from the proteins targeted by other viruses.

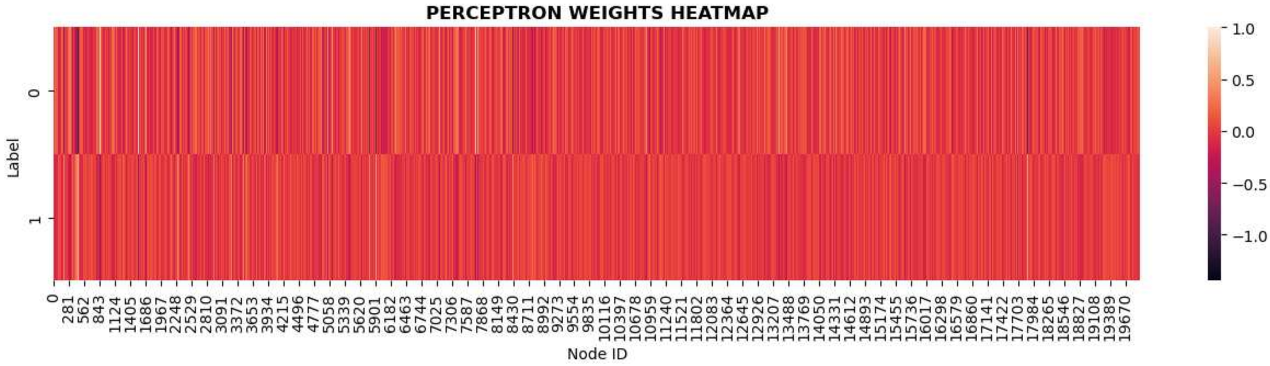


Figura 7.25: Caption

Another crucial finding is that, in each trial, samples belonging to $N1S$ are consistently predicted as having label 1, indicating the presence of an oncogenic virus layer.

Considering the impact of network dimension, we evaluated the model on a separate test set consisting of multilplexes with three layers corresponding to non-oncogenic viruses from the dataset. The fourth layer was a synthetic PPI network with 330 directly targeted proteins (matching Sars-Cov-2). These proteins were extracted from the dataset's virus targets.

The results, presented in the "SyntTest pred" column of Table 7.4, show values much closer to 0.5, which corresponds to random predictions, compared to those related to Sars-Cov-2. This suggests that the model's predictions for the synthetic network exhibit less bias and approach randomness. Hence, the unique characteristics of Sars-Cov-2 seem to have influenced the model's performance.

7.1.2 Weights

Given the model's simplicity, we can perform an analysis to interpret the obtained results. Specifically, the model has two sets of weights connecting specific input neurons to the two possible outputs. Since each input neuron is associated with a normalized page-rank versatility value, we can expect that if a protein's high relevance is linked to a specific output neuron, the corresponding weight will be large. Conversely, if the protein's importance is associated with the other output label, the associated weight is expected to be strongly negative.

We aim to identify proteins that are highly relevant in the classification process. To achieve this, one possible strategy is to calculate the difference between the weights associated with label 0 and label 1 for each input feature. The input features with the highest absolute differences were considered to be associated with proteins that are highly relevant for classification. In each trial, we selected the top 50 proteins based on this criterion.

	TL1	PV5	PV18	PV16	HV8P	HC1	HBC	EB	None
6 elements	15.1	29.9	18.0	12.0	11.9	11.9	12.6	11.9	24.8
5 elements	19.6	31.9	22.2	17.1	17.0	16.9	17.5	16.9	19.9
4 elements	24.5	34.3	26.8	22.6	22.6	22.4	22.9	22.4	15.4

Tabella 7.5: Caption

In order to obtain a robust set of proteins of interest, the idea is to intersect the results from different trials. By taking all the trials the resulting list of 7 proteins is the following:

CBX3,HIST1H2BG,HIST2H3A,MECP2,NCK2,NPSR1,TRIM29

To assess the generalization capabilities of the previously discussed models, it is important to consider appropriate subsets of the entire set of trials. A logical approach would be to exclude the trial with the lowest generalization capability, namely the $PV5$ trial.

To investigate this assumption, the mean values of the number of intersections were computed for all possible combinations of 4, 5, and 6 trials, starting from different initial sets. This involved considering all trials and all possible combinations of 7 out of the 8 available trials, with one combination excluding one trial at a time. The results, presented in Table 7.5, confirmed the initial hypothesis. Excluding the $PV5$ trial resulted in a significantly higher number of intersections compared to the other cases, indicating more similar results among the remaining trials.

The list of the intersection of the remaining 7 trials is composed by the following 29 proteins:

CBX3, CTBP1, DYRK1B, FBXO3, GATA4, HIPK2, HIST1H2BG, HIST2H3A, LZTS2, MECP2, MYB, MYBL1, NCK2, NKX2-1, NKX2-5, NLK, NPPA, NPSR1, PPM1D, RASSF5, RGMA, SENP1, SENP2, SKI, SP100, TP53INP1, TRIM29, ZBTB4

An alternative approach is to select the top 50 proteins with the highest weights for each of the two possible labels. This allows us to identify proteins that are more likely to be associated with either class.

Considering the weights associated with the presence of an oncogenic virus layer, the intersection of the results for the 8 trials yields the following proteins:

CBX3, DYRK1B, HDAC8, HIPK2, HIST1H2BG, HIST2H3A, MECP2, NCK2, NKX2-1, NLK, NPSR1, TAF6L, TRIM29

While by exluding the $PV5$ trials the intersection is:

CBX3, CHMP4B, CTBP1, DYRK1B, FBXO3, GATA4, HDAC8, HIPK2, HIST1H2BG, HIST2H3A, LZTS2, MECP2, MYB, MYBL1, NCK2, NKX2-1, NKX2-5, NLK, NPPA, NPSR1, PPM1D, RASSF5, RGMA, SENP1, SENP2, SKI, SP100, TAF6L, TP53INP1, TRIM29, ZBTB4

On the other hand, by considering the weights associated to the presence of only non-oncogenic virus layers, the result for the intersection of the results for the 8 trials and the one associated to the exclusion of the $PV5$ one, correspond:

ATP6AP2, BST1, CCNC, DEFA5, E2F3, EEF1A2, GOPC, HES1, HSPG2, IDE, IWS1, LYPD3, PDK1, RRM1, SNX17, TCF12, TNPO1, UBA2

7.2 Random Forest

The second model used for the classification task is the random forest. A Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. The second model used for the classification task is the random forest. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.

Trial Name	Excluded Onco Virus	Train Acc	Val Acc	OncoTest pred	SarsTest pred	SyntTest pred
EB	Epstein-Barr	1.000	0.949	0.974	0.720	0.820
HBC	Hepatitis B gen. C, ayr	1.000	0.947	0.838	0.920	0.780
HC1	Hepatitis C gen. 1a	1.000	0.939	0.906	0.660	0.490
HV8P	Hum. herpesvirus 8 type P	1.000	0.945	0.978	0.740	0.660
PV16	Hum. papillomavirus type 16	1.000	0.959	0.873	0.800	0.580
PV18	Hum. papillomavirus type 18	1.000	0.964	0.890	0.850	0.700
PV5	Hum. papillomavirus type 5	1.000	0.956	0.442	0.920	0.830
TL1	Hum. T-cell leukemia 1	1.000	0.963	0.890	0.740	0.830

Tabella 7.6: The table presents the performance results of individual Random Forest models trained using datasets in which samples containing specific oncogenic viruses were excluded. Each trained model was subsequently tested on different sample sets: OncoTest consists of samples containing the excluded oncogenic virus PPI layer, SarsTest contains samples with the SARS-CoV-2 PPI network, and SyntTest comprises synthetic layers composed of the nearest neighbors of 330 proteins randomly selected from the proteins targeted by other viruses.

In a Random Forest, each decision tree is built independently using a random subset of the training data and a random subset of the input features. This randomness introduces diversity among the trees, reducing the risk of overfitting and improving the overall accuracy and robustness of the model.

During prediction, each tree in the Random Forest independently makes a prediction, and the final prediction is determined by combining the individual predictions through voting or averaging. This ensemble approach helps to improve the model's generalization and reduce the impact of individual trees' biases.

Random Forests also provide measures of feature importance, indicating the relative contribution of each input feature in making accurate predictions, which aids in interpretability.

The dataset processing and generation of different trials for testing the model are the same as in the perceptron case. The parameters for the random forest are selected through optimization based on the accuracy on the validation set. The final set of parameters is:

n estimators	100
max depth	None
min samples split	2
min samples leaf	1
max features	sqrt
criterion	entropy

In Tab.7.6 are reported the same quantities presented in Tab.7.4 for the different trials, in order to be able to compare the two methods.

The results for the OncoTest pred column are slightly better compared to the perceptron method, suggesting a potentially stronger generalization capability. However, the values in the SarsTest pred column are smaller, indicating a lower confidence in identifying Sars-Cov-2 as an oncogenic virus. Additionally, the SyntTest pred values are closer to 1, implying a higher likelihood of classifying random viruses as oncogenic. This trend is opposite to that of the perceptron, where random viruses were less likely to be classified as oncogenic, while Sars-Cov-2 showed the opposite pattern.

7.2.1 Scores

The random forest framework provides a feature importance score for each entry in the model. This score is calculated using the entropy criterion during the training process, following these steps:

1. Multiple decision trees are trained using the random forest algorithm, with each tree built on a different subset of the training data.
2. For each tree, the information gain or decrease in entropy is computed at each split point.

3. The average information gain across all trees is calculated for each feature.
4. The feature importance scores are normalized to ensure they sum up to 1 or 100%.
5. Features are sorted based on their importance scores in descending order.
6. The importance of each feature is determined by its ranking, with higher scores indicating greater importance in predicting the target variable.

In each trial, the scores for each entry are computed, and the 50 proteins with the highest scores are considered the most representative in the classification task. The following proteins are listed as belonging to the top 50 most important ones in all the 8 trials.

CBX3, CTBP1, HIST1H2BG, HIST1H3A, NTRK1, SUMO2

Like in the perceptron case, it's possible to enlarge such a set by excluding from the intersection the results from the $PV5$ trial, which is the one that provides the worst generalization capabilities. The results corresponds to the following list of 16 proteins.

CBX3, CBX4, CTBP1, DYRK1B, HIPK2, HIST1H2BG, HIST1H3A, HIST2H2BE, MECP2, MYB, NKX2-1, NLK, NTRK1, SENP1, SP100, SUMO2

7.3 "Experimental" and "Co-Expression" Networks

The classification task using the perceptron was performed by using the "Experimental" and "Co-expression" protein-protein interaction (PPI) datasets derived from STRING. In the "Experimental" case, a cutoff of 700 non-zero entries was applied out of a total of 12,984 proteins in the human PPI network. For the "Co-expression" dataset, a cutoff of 500 non-zero entries was applied out of 4,499 proteins.

The procedure for constructing the different trials, excluding samples containing oncogenic viruses and reserving them for testing the model, followed the same methodology as described earlier for the BioSTRING dataset.

In the case of the "Experimental" dataset, each model was trained for 500 epochs, while for the "Co-expression" dataset, training was performed for 300 epochs, as these limits were found to result in plateaued accuracies.

Table 7.7 shows the results from the model training for the "Experimental" case, while Table 7.8 presents the results for the "Co-expression" case.

Trial Name	Excluded Onco Virus	Train Acc	Val Acc	OncoTest pred
EB	Epstein-Barr	0.999	0.996	0.789
HBC	Hepatitis B gen. C, ayr	0.996	0.996	0.817
HC1	Hepatitis C gen. 1a	0.995	0.998	0.678
HV8P	Hum. herpesvirus 8 type P	0.995	0.996	0.324
PV16	Hum. papillomavirus type 16	0.997	0.996	0.938
PV18	Hum. papillomavirus type 18	0.996	0.992	0.887
PV5	Hum. papillomavirus type 5	0.997	0.996	0.696
TL1	Hum. T-cell leukemia 1	0.998	0.992	0.958

Tabella 7.7: The table presents the performance results of individual perceptron models considering the "Experimental" dataset, trained using datasets in which samples containing specific oncogenic viruses were excluded. Each trained model was subsequently tested on samples containing the excluded oncogenic virus PPI layer (OncoTest).

The generalization capabilities of the models, tested through predictions on the OncoTest samples, were found to be weaker for both the "Experimental" and "Co-expression" datasets compared to the

Trial Name	Excluded Onco Virus	Train Acc	Val Acc	OncoTest pred
EB	Epstein-Barr	0.978	0.984	0.054
HBC	Hepatitis B gen. C, ayr	0.957	0.939	0.887
HC1	Hepatitis C gen. 1a	0.959	0.950	0.142
HV8P	Hum. herpesvirus 8 type P	0.951	0.925	0.019
PV16	Hum. papillomavirus type 16	0.959	0.955	0.876
PV18	Hum. papillomavirus type 18	0.962	0.959	0.557
PV5	Hum. papillomavirus type 5	0.968	0.984	0.120
TL1	Hum. T-cell leukemia 1	0.953	0.955	0.570

Tabella 7.8: The table presents the performance results of individual perceptron models considering the "Co-expression" dataset, trained using datasets in which samples containing specific oncogenic viruses were excluded. Each trained model was subsequently tested on samples containing the excluded oncogenic virus PPI layer (OncoTest).

BioSTRING case. However, while the performance for the "Experimental" dataset remained relatively good, the performance for the "Co-expression" dataset was consistently poor across the majority of the trials.

In both cases, the top 50 proteins with the highest absolute difference between weights associated with labels 0 and 1 ,were identified. For the "Experimental" dataset, the intersection of these protein lists was obtained by excluding the trials $HC1$, $HV8P$, and $PV5$, which exhibited the poorest classification performance on the OncoTest set. The resulting protein list includes:

SERPIND1, GNB3, GNB4, GRAP, KCNJ3, GNB2, MAPK7, MEOX1, PRDX6, CS, RNF2, UBR4, GGT5, MAPK4, SEPT9, LDB3, CSNK1A1

For the "Co-expression" case, the intersection was determined by considering only the trials HBC and $PV16$, which demonstrated the best predictions on the test set. The resulting protein list includes:

GNPNAT1, SPRTN, March3, SUPT6H, PPP2R1A, ATP13A2, CYLC1, PHF7, ATP13A4, ATP13A5, CORO2A, CORO6, MTMR8, DEFB123, DDX51, HM13, TENM1, CORO2B

Chapter 8

Discussion

The network theory-based approach utilized in the analysis yielded a set of findings that can be further interpreted in a biological context, aiming to explore potential connections between these results and existing knowledge regarding the relationship between viruses, tumors, and oncogenic mechanisms in general.

Subsequently, the focus will shift towards the main objective of the study, which is to provide insights into the tumor-related characteristics of SARS-CoV-2 based on the obtained results.

8.1 Perceptron highly relevant features

Once we have obtained the list of proteins that are highly relevant in the perceptron classification task, we can further analyze them and propose a biological interpretation.

Firstly, it is important to highlight that proteins with a high absolute difference between their corresponding weights should exhibit high centrality values in both the $N1O$ and N cases. This indicates that these proteins are highly relevant within the PPI networks in general.

To explore the biological significance of these proteins, we consider the intersection of all the trials except for the $PV5$ trial mentioned earlier. A graphical representation of this intersection, obtained from the STRING web page, is presented in Fig.8.26.

A functional enrichment analysis is performed over this set of genes with the following results:

Name	**pValue**	**Bonferroni**	**Overlap**
structural constituent of chromatin	3.634E-7	3.452E-5	4/106
protein dimerization activity	4.041E-5	3.839E-3	6/1251
protein tyrosine kinase activity	8.427E-5	8.006E-3	3/148
protein heterodimerization activity	9.936E-5	9.440E-3	4/438
protein serine/threonine kinase activity	1.242E-4	1.180E-2	4/464

Tabella 8.9: The table presents the most over-represented Gene Ontology Molecular Function terms related to the list of proteins highly relevant for the classification task performed with the perceptron algorithm. For each of these is reported the associated p-value, a the Bonferroni adjusted p-value, and the overlap, i.e. the ratio between the number of proteins in the list which are present in the term and the total number of proteins in the term.

8.1.1 Chromatin structure

Both the protein list and the result of the GO molecular function enrichment analysis highlight the relevance of histones and factors involved in chromatin structure.

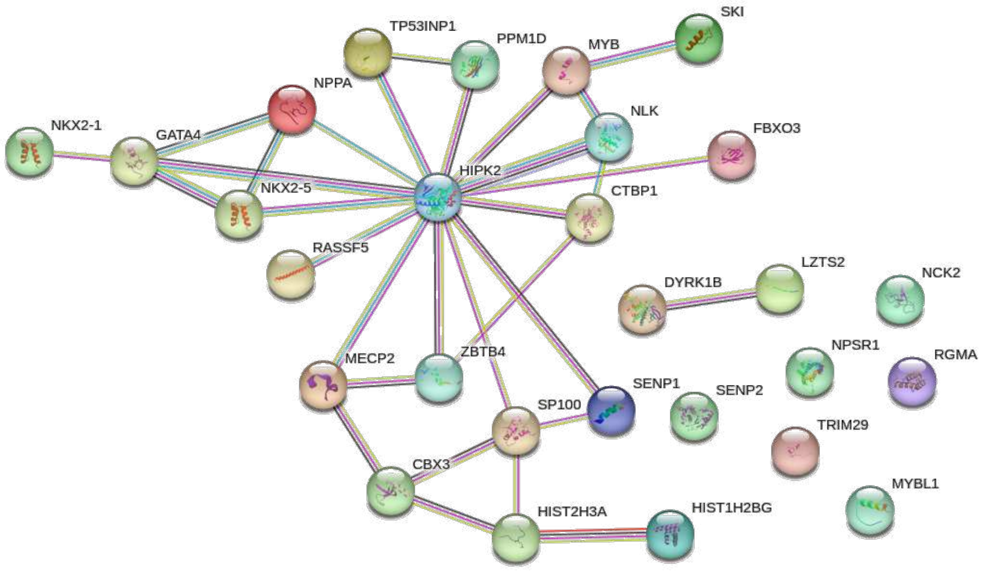


Figura 8.26: Representation of the PPI network based on the list of proteins highly relevant for the classification task performed with the perceptron. The structure is generated from the STRING web page, and it wasn't applied any cutoff to the links between the proteins `https://string-db.org/cgi/network?taskId=bZiJXAv9zPEv&sessionId=bnKXMWMsf0q5`

Name	pValue	Bonferroni	Overlap
REACTOME TRANSCRIPTIONAL REGULATION OF GRANULOPOIESIS	7.459E-10	3.677E-7	6/90
PID CMYB PATHWAY	4.265E-8	2.103E-5	5/84
REACTOME DNA METHYLATION	1.007E-6	4.964E-4	4/65
REACTOME SIRT1 NEGATIVELY REGULATES RRNA EXPRESSION	1.208E-6	5.958E-4	4/68
REACTOME SIGNALING BY WNT	1.771E-6	8.731E-4	6/331

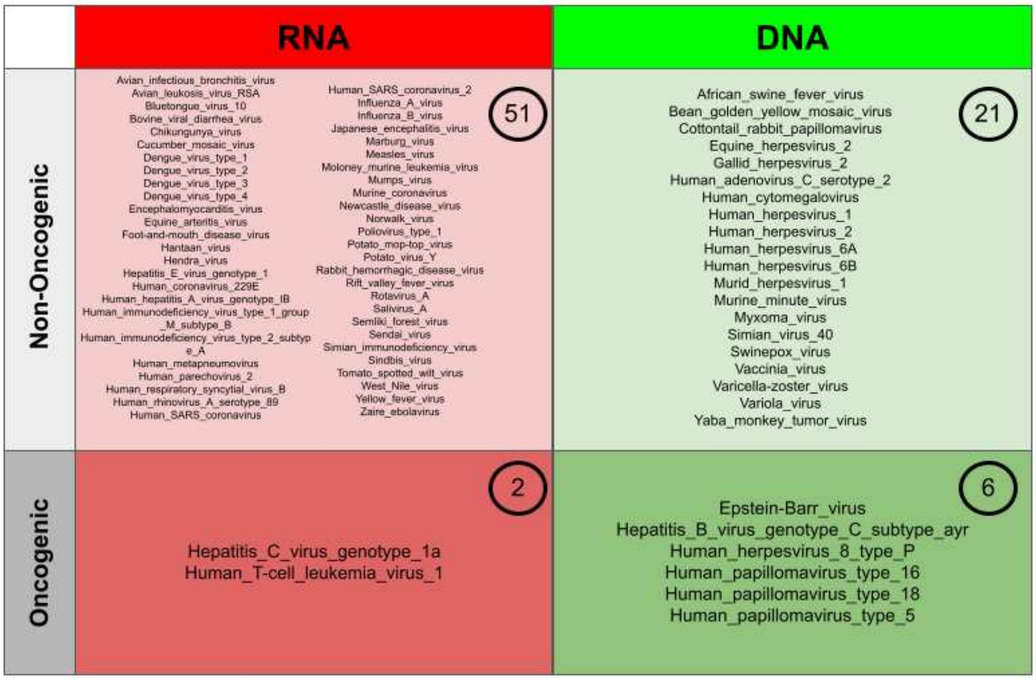
Tabella 8.10: The table presents the most over-represented Pathways terms related to the list of proteins highly relevant for the classification task performed with the perceptron algorithm. For each of these is reported the associated p-value, a the Bonferroni adjusted p-value, and the overlap, i.e. the ratio between the number of proteins in the list which are present in the term and the total number of proteins in the term.

Histones are a group of proteins that play a fundamental role in packaging and spatially organizing DNA within the nucleus of eukaryotic cells. They interact with DNA molecules, creating links between different positions on the DNA string. Histones are crucial for regulating gene expression, as they compact and condense DNA, enabling it to fit within the limited space of the nucleus. They are involved in essential processes such as DNA replication, repair, and chromosome segregation during cell division. Histones provide structural stability to chromosomes and help maintain the integrity of the genome.

Alterations in chromatin structure have been found to contribute to the development and progression of cancer[46]. Various mechanisms can disrupt normal chromatin organization in cancer cells, including changes in DNA methylation patterns, histone modifications, chromatin remodeling, and higher-order chromatin folding [47][48].

One key aspect is the epigenetic regulation of gene expression through modifications of chromatin structure. Epigenetic modifications, including DNA methylation and histone modifications, can lead to the silencing or activation of specific genes involved in cancer-related processes such as cell proliferation, differentiation, and metastasis.

In addition to the functional role of these specific proteins in tumor development, the difference in the



	RNA	DNA
Non-Oncogenic	51 Avian_infectious_bronchitis_virus Avian_leukosis_virus_RSA Bluetongue_virus_10 Bovine_viral_diarrhea_virus Chikungunya_virus Cucumber_mosaic_virus Dengue_virus_type_1 Dengue_virus_type_2 Dengue_virus_type_3 Dengue_virus_type_4 Encephalomyocarditis_virus Equine_arteritis_virus Foot-and-mouth_disease_virus Hantaan_virus Hendra_virus Hepatitis_E_virus_genotype_1 Human_coronavirus_229E Human_hepatitis_A_virus_genotype_IB Human_immunodeficiency_virus_type_1_group_M_subtype_B Human_immunodeficiency_virus_type_2_subtype_A Human_metapneumovirus Human_parechovirus_2 Human_respiratory_syncytial_virus_B Human_rhinovirus_A_serotype_89 Human_SARS_coronavirus Human_SARS_coronavirus_2 Influenza_A_virus Influenza_B_virus Japanese_encephalitis_virus Marburg_virus Measles_virus Moloney_murine_leukemia_virus Mumps_virus Murine_coronavirus Newcastle_disease_virus Norwalk_virus Poliovirus_type_1 Potato_mop-top_virus Potato_virus_Y Rabbit_hemorrhagic_disease_virus Rift_valley_fever_virus Rotavirus_A Salivirus_A Semliki_forest_virus Sendai_virus Simian_immunodeficiency_virus Sindbis_virus Tomato_spotted_wilt_virus West_Nile_virus Yellow_fever_virus Zaire_ebolavirus	21 African_swine_fever_virus Bean_golden_yellow_mosaic_virus Cottontail_rabbit_papillomavirus Equine_herpesvirus_2 Gallid_herpesvirus_2 Human_adenovirus_C_serotype_2 Human_cytomegalovirus Human_herpesvirus_1 Human_herpesvirus_2 Human_herpesvirus_6A Human_herpesvirus_6B Murid_herpesvirus_1 Murine_minute_virus Myxoma_virus Simian_virus_40 Swinepox_virus Vaccinia_virus Varicella-zoster_virus Variola_virus Yaba_monkey_tumor_virus
Oncogenic	2 Hepatitis_C_virus_genotype_1a Human_T-cell_leukemia_virus_1	6 Epstein-Barr_virus Hepatitis_B_virus_genotype_C_subtype_ayr Human_herpesvirus_8_type_P Human_papillomavirus_type_16 Human_papillomavirus_type_18 Human_papillomavirus_type_5

Figura 8.27: Classification of the viruses in the dataset based on their genetic material (DNA vs RNA) and the oncogenic feature. In the top right of each box is reported the total number of viruses with that specific set of features

nature of viruses may also contribute to the observed association. Viruses can be categorized as DNA viruses or RNA viruses. DNA viruses replicate their genetic material using host cellular machinery and DNA polymerases, primarily in the nucleus of the host cell. In contrast, RNA viruses replicate their genetic material using an RNA-dependent RNA polymerase (RdRp) and typically replicate in the cytoplasm of the host cell without accessing the nucleus.

Therefore, it is more likely that DNA viruses interact with histones, which are predominantly located in the nucleus of the cell. On the other hand, RNA viruses generally replicate in the cytoplasm, where histones are less abundant. However, it's important to note that certain RNA viruses, such as retroviruses, which convert their RNA genome into DNA intermediates, can interact with histones during the integration of viral DNA into the host cell's genome.

To further investigate this relationship, it is worth examining the distribution of viruses in the dataset between DNA and RNA viruses, including both oncogenic and non-oncogenic types. This analysis can reveal any imbalances or patterns. A schematic representation of the results is presented in Fig.8.27.

The number of DNA viruses is higher than RNA viruses in the oncogenic case, while in the non-oncogenic case, the trend is reversed. This discrepancy raises doubts about the classification task. The classifier may focus more on distinguishing between DNA and RNA viruses rather than between oncogenic and non-oncogenic viruses due to the imbalance in these classes.

Examining the performance of the perceptron and random forest classifiers (Tab.7.4, Tab.7.6), it is observed that the trial associated with the removal and testing of the *Human papillomavirus type 5* (a DNA virus) leads to the worst generalization performance.

This observation contradicts the hypothesis that the classification is based solely on virus type, as the performance should have significantly decreased when excluding and testing on an RNA virus. Furthermore, the training samples are even more imbalanced toward the DNA virus type, and the testing is performed with a virus of the opposite type.

Moreover, the performance across different trials, except for the $PV5$ case, shows minimal variance, and there is no clear distinction between cases where DNA or RNA viruses are excluded. This supports the hypothesis that proteins involved in chromatin structure may play a crucial role in distinguishing between the oncogenic and non-oncogenic cases.

Based on these considerations, it can be concluded that the classification model may not heavily rely on the type of RNA or DNA viruses. This provides support for the hypothesis that the model can identify oncogenic-related features beyond viral types.

8.1.2 SUMO related proteins

In the analysis of Reactome pathways, it is evident that there are several pathways related to SUMOylation. SUMO (Small Ubiquitin-like Modifier) is a group of proteins that play a crucial role in post-translational modifications and the regulation of various cellular processes in biology. Within the SUMOylation pathway, SUMO molecules covalently attach to target proteins, resulting in modifications to their function, localization, stability, and interactions with other molecules. Notably, the proteins list includes key enzymes involved in this process, such as SENP1 and SENP2. Numerous studies have explored the potential association between cancer development and the upregulation of SUMOylation mechanisms, which may provide an explanation for the significant relevance of these proteins in the analysis conducted in this study. [49, 50]

8.1.3 WNT signaling pathway and P53 inhibition

Another protein of significant interest is $HIPK2$, a protein kinase that plays a crucial role in various cellular processes, including cell proliferation, apoptosis, DNA damage response, and development. $HIPK2$ functions as a transcriptional co-regulator and regulates protein stability.

$HIPK2$ is particularly important in the regulation of $p53$, a tumor suppressor protein. It phosphorylates $p53$, leading to its activation and stabilization, which promotes cell cycle arrest and apoptosis in response to DNA damage. This interaction between $HIPK2$ and $p53$ helps maintain genomic stability and prevents the formation of cancer cells.

$HIPK2$'s high relevance is further supported by its connections to many other proteins in the analyzed list, indicating its role as a hub for highly relevant proteins in the classification task.

The absence of $p53$ (or $TP53$) in the protein list is noteworthy. $p53$ is a well-known protein in cancer research as its dysregulation can disrupt essential cellular processes and contribute to tumor development. Its absence from the list may be because it acts as a hub in the human protein-protein interaction network and is highly relevant in both normal and oncogenic samples, providing limited discriminatory information for the classification task. However, proteins associated with $p53$ and its related biological processes, such as $HIPK2$ and $TP53INP1$ (a protein that positively regulates $TP53$ and $TP73$ and stimulates their capacity to induce apoptosis and regulate cell cycle), reinforce the notion that the most relevant features for classification are indeed related to mechanisms involved in tumor development.

8.2 Random Forest high relevant features

Moreover, a similar analysis can be applied to the results of the feature importance analysis on the Random Forest model. In the following, we will consider the list of the proteins derived from the intersection of the top 50 proteins for the 7 trials excluding the $PV5$ one. Figure 8.28 presents the graphical representation of the considered list of proteins, while Table 8.11 shows the results of the enrichment analysis considering the Reactome pathways.

It is noteworthy that most of these pathways are either the same or related to the ones associated with the perceptron list in Table 8.10.

For instance, the pathway related to the sumoylation process stands out as the most overrepresented one. The WNT signaling pathway also holds significant importance, with the inclusion of $HIPK2$ in this set.

Name	**pValue**	**Bonferroni**	**Overlap**
REACTOME SUMOYLATION	2.194E-8	6.692E-6	6/169
REACTOME TRANSCRIPTIONAL REGULATION OF GRANULOPOIESIS	2.866E-6	8.741E-4	4/90
REACTOME RNA POLYMERASE I PROMOTER ESCA-PE	2.996E-6	9.137E-4	4/91
REACTOME SIRT1 NEGATIVELY REGULATES RRNA EXPRESSION	1.208E-6	5.958E-4	4/68
REACTOME POSITIVE EPIGENETIC REGULATION OF RRNA EXPRESSION	5.510E-6	1.681E-3	4/106
REACTOME SIGNALING BY WNT	2.668E-5	8.138E-3	5/331

Tabella 8.11: The table presents the most over-represented Pathways terms related to the list of proteins highly relevant for the classification task performed with the Random Forest algorithm. For each of these is reported the associated p-value, a the Bonferroni adjusted p-value, and the overlap, i.e. the ratio between the number of proteins in the list which are present in the term and the total number of proteins in the term.

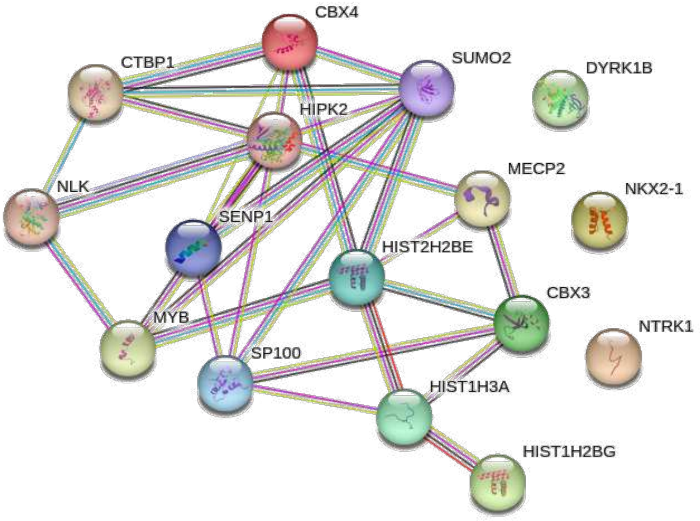


Figura 8.28: Representation of the PPI network based on the list of proteins highly relevant for the classification task performed with the Random Forest algorithm. The structure is generated from the STRING web page, and it wasn't applied any cutoff to the links between the proteins `https://string-db.org/cgi/network?taskId=bZiJXAv9zPEv&sessionId=bnKXMWMsf0q5`

Of particular significance is the presence of three histones, namely $HIST1H2BG, HIST1H3A, HIST2H2BE$, which is one more than in the perceptron case. This further reinforces the notion of the critical role that chromatin structure manipulation can play in the oncogenic processes associated with virus infection.

8.3 LIC intersection of oncogenic multilayer networks

The investigation of the protein list, as described in Chapter 5.1.2, derived from the intersection of all the LICs extracted from the samples belonging to the O combination set, provides interesting insights. A graphical representation of the protein interactions, generated using STRING, is presented in Figure 8.29.

As expected, most of the proteins exhibit connections with each other, reflecting their involvement in the PPI networks of various oncogenic viruses. The isolated proteins are likely connected to the largest component through links obtained from the BIOGRID database, as explained in Chapter 3.2.3.

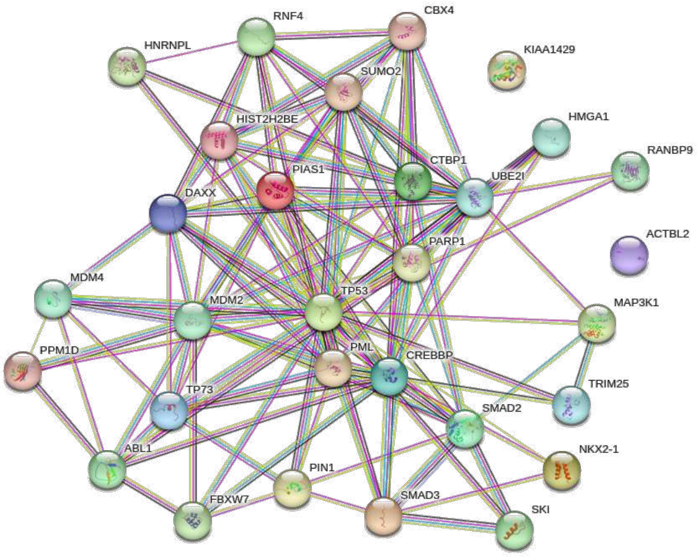


Figura 8.29: Representation of the PPI network based on the list of proteins that form a component in all the PPI networks associated to oncogenic viruses. The structure is generated from the STRING web page, and it wasn't applied any cutoff to the links between the proteins `https://string-db.org/cgi/network?taskId=bZiJXAv9zPEv&sessionId=bnKXMWMsf0q5`

Notably, the $TP53$ protein assumes a central role as a hub, which aligns with expectations.

Additionally, the presence of proteins such as histone $HIST2HBG$, SKI (involved in chromatin structure), and important proteins related to sumoylation, such as $SUMO2$ and $CBX4$, is worth highlighting.

To gain further insights into the nature of this protein set, a functional enrichment analysis was conducted. This analysis involved examining the GO molecular (Tab.8.12) functions and biological pathways (Tab.8.13) associated with these proteins.

Name	**pValue**	**Bonferroni**	**Overlap**
DNA-binding transcription factor binding	1.165E-15	3.087E-13	15/595
SUMO transferase activity	8.266E-14	2.190E-11	7/37
transcription coregulator activity	1.430E-11	3.789E-9	12/569
chromatin binding	6.050E-8	1.603E-5	10/726
p53 binding	1.001E-7	2.654E-5	5/77

Tabella 8.12: The table presents the most over-represented Gene Ontology Molecular Function terms related to the list of proteins that form a component in all the PPI networks associated to oncogenic viruses. For each of these is reported the associated p-value, a the Bonferroni adjusted p-value, and the overlap, i.e. the ratio between the number of proteins in the list which are present in the term and the total number of proteins in the term.

Both tables highlight the presence of relevant terms that align with the aspects analyzed in the classification tasks using both the perceptron and random forest algorithms.

The pathways associated with cancer and the regulation of $TP53$ are evident. However, the most significant importance is attributed to pathways related to sumoylation, which further supports the findings discussed in the previous section. Additionally, terms associated with the WNT signaling pathway and chromatin structure are present, reinforcing the conclusions drawn earlier.

Name	**pValue**	**Bonferroni**	**Overlap**
REACTOME SUMOYLATION	5.705E-14	5.317E-11	11/187
PID P53 REGULATION PATHWAY	2.665E-13	2.483E-10	8/59
KEGG PATHWAYS IN CANCER	1.287E-8	1.199E-5	9/325
REACTOME REGULATION OF TP53 ACTIVITY	2.559E-8	2.385E-5	7/156
KEGG WNT SIGNALING PATHWAY	1.390E-5	5.422E-4	5/151

Tabella 8.13: The table presents the most over-represented Pathways terms related to the list of proteins that form a component in all the PPI networks associated to oncogenic viruses. For each of these is reported the associated p-value, a the Bonferroni adjusted p-value, and the overlap, i.e. the ratio between the number of proteins in the list which are present in the term and the total number of proteins in the term.

8.4 "Experimental" and "Co-expression" dataset perceptron classifications

It's worth mentioning also some biological insights of the list of most relevant proteins coming form the perceptron classification task using the "Experimental" and "Co-expression" datasets.

In the "Experimental" case, the list of most relevant proteins includes $MAPK4$ and $MAPK7$, which are involved in crucial cellular processes such as proliferation, differentiation, and cell survival, all of which are relevant in the context of tumor cells. The enrichment analysis reveals overrepresented pathways related to G proteins, which are molecular switches involved in transmitting signals from external stimuli to the interior of a cell. While it is reasonable to assume that viruses interact with the cell's outer regions, these pathways do not seem to be directly linked to the differences between oncogenic and non-oncogenic viruses.

On the other hand, the "Co-expression" analysis shows a scattered network of proteins, with many nodes remaining unconnected to others. The functional enrichment analysis does not yield meaningful results, describing generic cellular processes or those not easily associated with cancer development.

Based on these results, it can be concluded that in order to obtain an informative classification outcome, a larger network is preferable, even if it may contain biases in the experimental data. Such a network provides a richer description of the biological system, from which important information can be extracted using appropriate methods. In contrast, smaller networks may have fewer biases but are more likely to produce false negatives, limiting the model's ability to provide a satisfactory description of the underlying system.

8.5 Sars-Cov2 classification

After establishing the classification models and generating predictions for multiplexes that include Sars-Cov2 layers, it becomes possible to determine the confidence level at which Sars-Cov2 can be classified in the same class as the 8 known oncogenic viruses.

In fact the objective of the algorithms is to classify samples from the N class (non-oncogenic) and samples from the $N1O$ class (containing an oncogenic virus layer). Since the distinction between these samples lies in the presence or absence of an oncogenic virus layer, the question can be reformulated as determining whether any of the layers in the multilayer network sample are associated with an oncogenic virus (label 1) or if all layers belong to the non-oncogenic class (label 0).

In order to do so it's possible to exploit the Bayes theorem. In the following we will refer to samples containing an oncogenic virus layer as 1, to the ones which doesn't contain it as 0. The probability of a given sample to have true label 1 given the fact that it was predicted as belonging to 1 is:

$$P(1^T|1^P) = \frac{P(1^P|1^T)\cdot P(1^T)}{P(1^P)} = \frac{P(1^P|1^T)\cdot P(1^T)}{P(1^P|1^T)\cdot P(1^T) + P(1^P|0^T)\cdot P(0^T)} \tag{8.18}$$

	Perceptron		**Random Forest**	
Trial	**Sens(1^P)**	**Spec(1^P)**	**Sens(1^P)**	**Spec(1^P)**
EB	0.932	0.933	0.932	0.965
HBC	0.925	0.916	0.925	0.974
HC1	0.916	0.923	0.939	0.947
HV8P	0.907	0.893	0.903	0.952
PV16	0.922	0.899	0.953	0.945
PV18	0.919	0.907	0.956	0.968
PV5	0.972	0.900	0.952	0.949
TL1	0.928	0.918	0.936	0.972

Tabella 8.14: Values of sensitivity and specificity from the validation confusion matrix from the model belonging to all the trials performed both with the perceptron and Random forest classification algorithms.

1^T refers to the samples actually belonging to $N1O$, 1^P to the samples which are predicted to belong to this class, and 0^T the samples actually belonging to N.

$P(1^P|1^T)$ corresponds to the sensitivity, i.e. the probability that a sample from $N1O$ is predicted as label 1. This information can be extracted from the confusion matrices of the validation set of each model as the fraction of samples with label 1, classified as 1.

$P(1^T)$ represents the prior probability associated with label 1, which signifies the prediction of a sample belonging to the N or $N1O$ combination set. It is possible to set the prior probability $P(1^T)$ as the fraction of oncogenic viruses divided by the total number of viruses in the dataset, yielding a value of 0.1 (i.e., 8 oncogenic viruses out of 80 total viruses). Consequently, $P(0^T)$ is calculated as $1 - P(1^T)$, resulting in 0.9 (i.e., 72 non-oncogenic viruses out of 80 total viruses).

Following the training of each trial of the model, the performance of predictions was evaluated on a test set comprising multiplexes with the Sars-Cov2 PPI as one of the layers. This test set consisted of 100 samples, and each sample was assigned a prediction of either class 1 or class 0.

The Bayesian formula presented before should be updated in order to take into account multiple predictions, and this is done in the following way:

$$P(1^T|P_1, P_2, ..., P_n) = \frac{\prod_{i=1}^{n} Sens(P_i) \cdot Prior}{\prod_{i=1}^{n} Sens(P_i) \cdot Prior + \prod_{i=1}^{n}(1 - Spec(P_i)) \cdot (1 - Prior)} \tag{8.19}$$

$P_1, P - 2, ..., P_n$ are multiples prediction results, each of which can assume values 1^P or 0^P. On the other hand we have the values of sensitivity and specificity associated for each possible prediction result. In particular $Sens(1^P) = P(1^P|1^T)$, $Sens(0^P) = 1 - Sens(1^P)$ and $Spec(1^P) = P(0^P|0^T)$, $Spec(0^P) = 1 - Spec(1^P)$. All these values can be easily extracted from the confusion matrices, and the values are reported in Tab.8.14.

In this way it's possible to compute the confidence at which Sars-Cov2 is predicted to be in the oncogenic viruses class.

By utilizing models from all the trials, employing both the perceptron and random forest frameworks, and evaluating 100 samples from the $N1S$ combination set, the results consistently indicate a high level of confidence in predicting Sars-Cov2 as an oncogenic virus. In fact, the predicted probabilities for Sars-Cov2 belonging to the oncogenic class are consistently close to 1, further reinforcing the notion of its classification as an oncogenic virus.

Chapter 9

Conclusion

This study presents a method for analyzing PPI networks in relation to virus infections and proposes a classification approach to distinguish between oncogenic and non-oncogenic viruses. The utilization of complex network theory and multilayer networks offers promising opportunities for extracting valuable information, although further exploration is needed.

The analysis of single-layer networks reveals the presence of biases in the distribution of certain features among different PPI networks. However, the implications of these biases were not thoroughly investigated in this study and warrant further consideration.

By leveraging the multilayer framework and examining topological features, meaningful quantities for distinguishing between oncogenic and non-oncogenic viruses can be identified. However, the classification of Sars-Cov2 based on different quantities yields varying conclusions, sometimes associating it with oncogenic viruses and sometimes with non-oncogenic viruses.

When combining the topological analysis findings with SVM classification, the results appear to be robust. However, additional benchmarks and checks are necessary to validate their consistency, particularly regarding the classification of Sars-Cov2. A comprehensive evaluation is required before drawing definitive conclusions.

Moreover, the machine learning models developed in this study used datasets primarily from STRING and, to a lesser extent, BIOGRID. Exploring alternative data sources and methodologies to construct more accurate and less biased PPI networks for both human and virus-host interactions could be a valuable avenue for future research.

While the study examined certain factors that could introduce biases in classification, such as network size and the distinction between RNA and DNA viruses, further investigation is warranted to explore other potential influences and enhance confidence in the classification procedure. This would necessitate interdisciplinary collaboration across virology, immunology, biology, and related fields to gain deeper insights into the mechanisms by which oncogenic viruses contribute to cancer development.

Overall, this study proposes a promising method for classifying oncogenic and non-oncogenic viruses, with the potential for future developments to increase its robustness and provide valuable biological insights for cancer research.

Regarding the class prediction of SARS-CoV-2, the overall results suggest a strong likelihood of shared characteristics between this virus and other oncogenic viruses. This aligns with existing research indicating a potential link between long COVID-19 symptoms and an increased risk of tumor development [1, 51]. However, given the complexity of the subject, drawing a final conclusion requires further comprehensive analysis, and every result presented in this work should undergo rigorous checks and evaluations.

Bibliography

[1] Hoai-Nga Thi Nguyen, Marie Kawahara, Cat-Khanh Vuong, Mizuho Fukushige, Toshiharu Yamashita, and Osamu Ohneda. SARS-CoV-2 m protein facilitates malignant transformation of breast cancer cells. *Frontiers in Oncology*, 12, June 2022. doi: 10.3389/fonc.2022.923467. URL `https://doi.org/10.3389/fonc.2022.923467`.

[2] Alberto Gómez-Carballa, Federico Martinón-Torres, and Antonio Salas. Is SARS-CoV-2 an oncogenic virus? *Journal of Infection*, 85(5):573–607, November 2022. doi: 10.1016/j.jinf.2022.08.005. URL `https://doi.org/10.1016/j.jinf.2022.08.005`.

[3] Michele Costanzo, Maria Anna Rachele De Giglio, and Giovanni Nicola Roviello. Deciphering the relationship between SARS-CoV-2 and cancer. *International Journal of Molecular Sciences*, 24(9): 7803, April 2023. doi: 10.3390/ijms24097803. URL `https://doi.org/10.3390/ijms24097803`.

[4] John T. Schiller and Douglas R. Lowy. Virus infection and human cancer: An overview. In *Viruses and Human Cancer*, pages 1–10. Springer Berlin Heidelberg, September 2013. doi: 10.1007/978-3-642-38965-8_1. URL `https://doi.org/10.1007/978-3-642-38965-8_1`.

[5] Matthew P. Thompson and Razelle Kurzrock. Epstein-barr virus and cancer. *Clinical Cancer Research*, 10(3):803–821, February 2004. doi: 10.1158/1078-0432.ccr-0670-3. URL `https://doi.org/10.1158/1078-0432.ccr-0670-3`.

[6] Nathan A. Krump and Jianxin You. Molecular mechanisms of viral oncogenesis in humans. *Nature Reviews Microbiology*, 16(11):684–698, August 2018. doi: 10.1038/s41579-018-0064-6. URL `https://doi.org/10.1038/s41579-018-0064-6`.

[7] Harald zur Hausen and Ethel-Michele de Villiers. Cancer "causation" by infections—individual contributions and synergistic networks. *Seminars in Oncology*, 41(6):860–875, December 2014. doi: 10.1053/j.seminoncol.2014.10.003. URL `https://doi.org/10.1053/j.seminoncol.2014.10.003`.

[8] Monique Schuijer and Els M.J.J. Berns. TP53 and ovarian cancer. *Human Mutation*, 21(3): 285–291, February 2003. doi: 10.1002/humu.10181. URL `https://doi.org/10.1002/humu.10181`.

[9] Anne-Lise Børresen-Dale. Tp53 and breast cancer. *Human Mutation*, 21(3):292–300, February 2003. doi: 10.1002/humu.10174. URL `https://doi.org/10.1002/humu.10174`.

[10] E Hickman. The role of p53 and pRB in apoptosis and cancer. *Current Opinion in Genetics &amp Development*, 12(1):60–66, February 2002. doi: 10.1016/s0959-437x(01)00265-9. URL `https://doi.org/10.1016/s0959-437x(01)00265-9`.

[11] Riccardo Di Fiore, Antonella D'Anneo, Giovanni Tesoriere, and Renza Vento. RB1 in cancer: Different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *Journal of Cellular Physiology*, 228(8):1676–1687, April 2013. doi: 10.1002/jcp.24329. URL `https://doi.org/10.1002/jcp.24329`.

[12] Kristian G. Andersen, Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. The proximal origin of SARS-CoV-2. *Nature Medicine*, 26(4):450–452, March 2020. doi: 10.1038/s41591-020-0820-9. URL `https://doi.org/10.1038/s41591-020-0820-9`.

[13] Chang Song, Zesong Li, Chen Li, Meiying Huang, Jianhong Liu, Qiuping Fang, Zitong Cao, Lin Zhang, Pengbo Gao, Wendi Nie, Xueyao Luo, Jianhao Kang, Shimin Xie, Jianxin Lyu, and Xiao Zhu. SARS-CoV-2: The monster causes COVID-19. *Frontiers in Cellular and Infection Microbiology*, 12, February 2022. doi: 10.3389/fcimb.2022.835750. URL `https://doi.org/10.3389/fcimb.2022.835750`.

[14] 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, November 2020. doi: 10.1089/nsm.2020.0011. URL `https://doi.org/10.1089/nsm.2020.0011`.

[15] Plamen Ch Ivanov, Kang K L Liu, and Ronny P Bartsch. Focus on the emerging new fields of network physiology and network medicine. *New Journal of Physics*, 18(10):100201, October 2016. doi: 10.1088/1367-2630/18/10/100201. URL `https://doi.org/10.1088/1367-2630/18/10/100201`.

[16] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, December 2010. doi: 10.1038/nrg2918. URL `https://doi.org/10.1038/nrg2918`.

[17] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5(1), June 2014. doi: 10.1038/ncomms5212. URL `https://doi.org/10.1038/ncomms5212`.

[18] Abhijeet R. Sonawane, Scott T. Weiss, Kimberly Glass, and Amitabh Sharma. Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10, April 2019. doi: 10.3389/fgene.2019.00294. URL `https://doi.org/10.3389/fgene.2019.00294`.

[19] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabasi. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601–1257601, February 2015. doi: 10.1126/science.1257601. URL `https://doi.org/10.1126/science.1257601`.

[20] Feixiong Cheng, István A. Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. *Nature Communications*, 10(1), March 2019. doi: 10.1038/s41467-019-09186-x. URL `https://doi.org/10.1038/s41467-019-09186-x`.

[21] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3:041022, Dec 2013. doi: 10.1103/PhysRevX.3.041022. URL `https://link.aps.org/doi/10.1103/PhysRevX.3.041022`.

[22] Liang Yu, Dandan Zhou, Lin Gao, and Yunhong Zha. Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods*, 192:85–92, August 2021. doi: 10.1016/j.ymeth.2020.08.006. URL `https://doi.org/10.1016/j.ymeth.2020.08.006`.

[23] Zaynab Hammoud and Frank Kramer. Multilayer networks: aspects, implementations, and application in biomedicine. *Big Data Analytics*, 5(1), July 2020. doi: 10.1186/s41044-020-00046-0. URL `https://doi.org/10.1186/s41044-020-00046-0`.

[24] Min Li, Yu Lu, Jianxin Wang, Fang-Xiang Wu, and Yi Pan. A topology potential-based method for identifying essential proteins from PPI networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(2):372–383, March 2015. doi: 10.1109/tcbb.2014.2361350. URL `https://doi.org/10.1109/tcbb.2014.2361350`.

[25] B. Chen, W. Fan, J. Liu, and F.-X. Wu. Identifying protein complexes and functional modules–from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, 15(2):177–194, June 2013. doi: 10.1093/bib/bbt039. URL https://doi.org/10.1093/bib/bbt039.

[26] Xianyi Lian, Xiaodi Yang, Shiping Yang, and Ziding Zhang. Current status and future perspectives of computational studies on human-virus protein-protein interactions. *Briefings in Bioinformatics*, 22(5), March 2021. doi: 10.1093/bib/bbab029. URL https://doi.org/10.1093/bib/bbab029.

[27] Qurat ul Ain Farooq, Zeeshan Shaukat, Sara Aiman, and Chun-Hua Li. Protein-protein interactions: Methods, databases, and applications in virus-host study. *World Journal of Virology*, 10(6):288–300, November 2021. doi: 10.5501/wjv.v10.i6.288. URL https://doi.org/10.5501/wjv.v10.i6.288.

[28] Nantao Zheng, Kairou Wang, Weihua Zhan, and Lei Deng. Targeting virus-host protein interactions: Feature extraction and machine learning approaches. *Current Drug Metabolism*, 20 (3):177–184, May 2019. doi: 10.2174/1389200219666180829121038. URL https://doi.org/10.2174/1389200219666180829121038.

[29] Arda Halu, Manlio De Domenico, Alex Arenas, and Amitabh Sharma. The multiplex network of human diseases. *npj Systems Biology and Applications*, 5(1), April 2019. doi: 10.1038/s41540-019-0092-5. URL https://doi.org/10.1038/s41540-019-0092-5.

[30] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford; New York, 2010. ISBN 9780199206650 0199206651. URL http://www.amazon.com/Networks-An-Introduction-Mark-Newman/dp/0199206651/ref=sr_1_5?ie=UTF8&qid=1352896678&sr=8-5&keywords=complex+networks.

[31] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, oct 1999. doi: 10.1126/science.286.5439.509. URL https://doi.org/10.1126%2Fscience.286.5439.509.

[32] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, 63(6), May 2001. doi: 10.1103/physreve.63.062101. URL https://doi.org/10.1103/physreve.63.062101.

[33] Pavel Berkhin. A survey on PageRank computing. *Internet Mathematics*, 2(1):73–120, January 2005. doi: 10.1080/15427951.2005.10129098. URL https://doi.org/10.1080/15427951.2005.10129098.

[34] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. doi: 10.1073/pnas.122653799. URL https://doi.org/10.1073/pnas.122653799.

[35] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. 2008. doi: 10.48550/ARXIV.0803.0476. URL https://arxiv.org/abs/0803.0476.

[36] Tiago P. Peixoto. Bayesian stochastic blockmodeling, nov 2019. URL https://doi.org/10.1002%2F9781119483298.ch11.

[37] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), jan 2011. doi: 10.1103/physreve.83.016107. URL https://doi.org/10.1103%2Fphysreve.83.016107.

[38] Manlio De Domenico. *Multilayer Networks: Analysis and Visualization*. Springer International Publishing, 2022. doi: 10.1007/978-3-030-75718-2. URL https://doi.org/10.1007/978-3-030-75718-2.

[39] Manlio De Domenico. Multilayer networks illustrated. 2022. doi: 10.17605/OSF.IO/GY53K. URL `https://osf.io/gy53k/`.

[40] Tiago P. Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4), October 2015. doi: 10.1103/physreve.92.042807. URL `https://doi.org/10.1103/physreve.92.042807`.

[41] Gabriel Valiente. The landscape of virus-host protein–protein interaction databases. *Frontiers in Microbiology*, 13, July 2022. doi: 10.3389/fmicb.2022.827742. URL `https://doi.org/10.3389/fmicb.2022.827742`.

[42] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, October 2014. doi: 10.1093/nar/gku1003. URL `https://doi.org/10.1093/nar/gku1003`.

[43] H.V. Cook, N.T. Doncheva, D. Szklarczyk, C. von Mering, and L.J. Jensen. Viruses.string: A virus-host protein-protein interaction database. *Viruses*, 10, Sep 2018. doi: 10.3390/v10100519.

[44] Arsham Ghavasieh, Sebastiano Bontorin, Oriol Artime, Nina Verstraete, and Manlio De Domenico. Multiscale statistical physics of the pan-viral interactome unravels the systemic nature of SARS-CoV-2 infections. *Communications Physics*, 4(1), April 2021. doi: 10.1038/s42005-021-00582-8. URL `https://doi.org/10.1038/s42005-021-00582-8`.

[45] R. Oughtred, J. Rust, C. Chang, and al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30:187–200, 2021. doi: 10.1002/pro.3978. URL `https://doi.org/10.1002/pro.3978`.

[46] Meng Wang, Benjamin D. Sunkel, William C. Ray, and Benjamin Z. Stanton. Chromatin structure in cancer. *BMC Molecular and Cell Biology*, 23(1), July 2022. doi: 10.1186/s12860-022-00433-6. URL `https://doi.org/10.1186/s12860-022-00433-6`.

[47] Shuai Zhao, C. David Allis, and Gang Greg Wang. The language of chromatin modification in human cancers. *Nature Reviews Cancer*, 21(7):413–430, May 2021. doi: 10.1038/s41568-021-00357-x. URL `https://doi.org/10.1038/s41568-021-00357-x`.

[48] Malcolm V. Brock, James G. Herman, and Stephen B. Baylin. Cancer as a manifestation of aberrant chromatin structure. *The Cancer Journal*, 13(1):3–8, January 2007. doi: 10.1097/ppo.0b013e31803c5415. URL `https://doi.org/10.1097/ppo.0b013e31803c5415`.

[49] Jacob-Sebastian Seeler and Anne Dejean. SUMO and the robustness of cancer. *Nature Reviews Cancer*, 17(3):184–197, January 2017. doi: 10.1038/nrc.2016.143. URL `https://doi.org/10.1038/nrc.2016.143`.

[50] Antti Kukkula, Veera K. Ojala, Lourdes M. Mendez, Lea Sistonen, Klaus Elenius, and Maria Sundvall. Therapeutic potential of targeting the SUMO pathway in cancer. *Cancers*, 13(17):4402, August 2021. doi: 10.3390/cancers13174402. URL `https://doi.org/10.3390/cancers13174402`.

[51] Daniela Milani, Lorenzo Caruso, Enrico Zauli, Adi Mohammed Al Owaifeer, Paola Secchiero, Giorgio Zauli, Donato Gemmati, and Veronica Tisato. p53/NF-kB balance in SARS-CoV-2 infection: From OMICs, genomics and pharmacogenomics insights to tailored therapeutic perspectives (COVIDomics). *Frontiers in Pharmacology*, 13, May 2022. doi: 10.3389/fphar.2022.871583. URL `https://doi.org/10.3389/fphar.2022.871583`.