Physiological modeling and biomechanical insights of cytokine modulation in COVID-19 and pulmonary fibrosis through a multi-dimensional graph-based approach

Rajarshi Ray[1#], Ratul Bhowmik[1#], Sagar Singh Shyamal[2#], Ajay Manaithiya[1#], Wenping Gong[3], Seppo Parkkila[1,4], Ashok Aspatwar[1*]

[1]Faculty of Medicine and Health Technology, Tampere University, FI-33014 Tampere, Finland

[2]Department of Pharmaceutical Engineering and Technology, Indian Institute of Technology (Banaras Hindu University), Varanasi, India

[3]Army Tuberculosis Prevention and Control Key Laboratory, Beijing Key Laboratory of New Techniques of Tuberculosis Diagnosis and Treatment, Institute for Tuberculosis Research, the 8th Medical Center of Chinese PLA General Hospital, Beijing, China.

[4]Fimlab Ltd, Tampere University Hospital, FI-33520 Tampere, Finland.

Abstract

COVID-19 has spread extensively worldwide, infecting millions and causing several fatalities because of its high severity. The illness was attributed to the SARS-CoV-2 virus, which specifically targets the respiratory and cardiovascular systems. Cytokine Release Syndrome (CRS) is a primary contributor to the tissue damage caused by COVID-19. CRS leads to abnormal release of cytokines and immune cells, which further disrupts cytokine activity, resulting in alterations in immune protein health-related biological pathways and accumulating aberrant levels of inflammatory proteins at infection sites that ultimately damage cells and tissues. Furthermore, COVID-19 patients with a previous history of respiratory difficulties develop pulmonary fibrosis (PF), resulting in atypical expression of fibrotic proteins and the release of cytokines. The presence of these inflammatory modulations in the lung network and subsequent respiratory failure often lead to medical emergencies. To understand the molecular-level alterations of the trisection of COVID-PF-CRS condition for identifying potential biomarkers, we established a series of knowledge graph-based hierarchical clustering matrix representations that mechanistically correlate the similarity of gene deregulated based on the commonality of pathways and phenotypes altered and modulated. We also elucidated the mechanism of action of toxicological and therapeutic chemicals using a multi-dimensional graph-based algorithm by understanding their modulation at the genetic level through pathways and phenotype alterations. Additionally, we also developed robust quantitative structural activity relation (QSAR) models in combination with system biology approaches against the top expressed hub genes of our knowledge graph to understand the key pharmacophores of biologically active small molecules that can be further implemented to develop drugs against COVID-19.

Keywords: Covid-19, SARS-CoV-2, Cytokine Release Syndrome, Pulmonary Fibrosis, inflammatory responses, system biology, knowledge graph, hierarchical clustering, QSAR

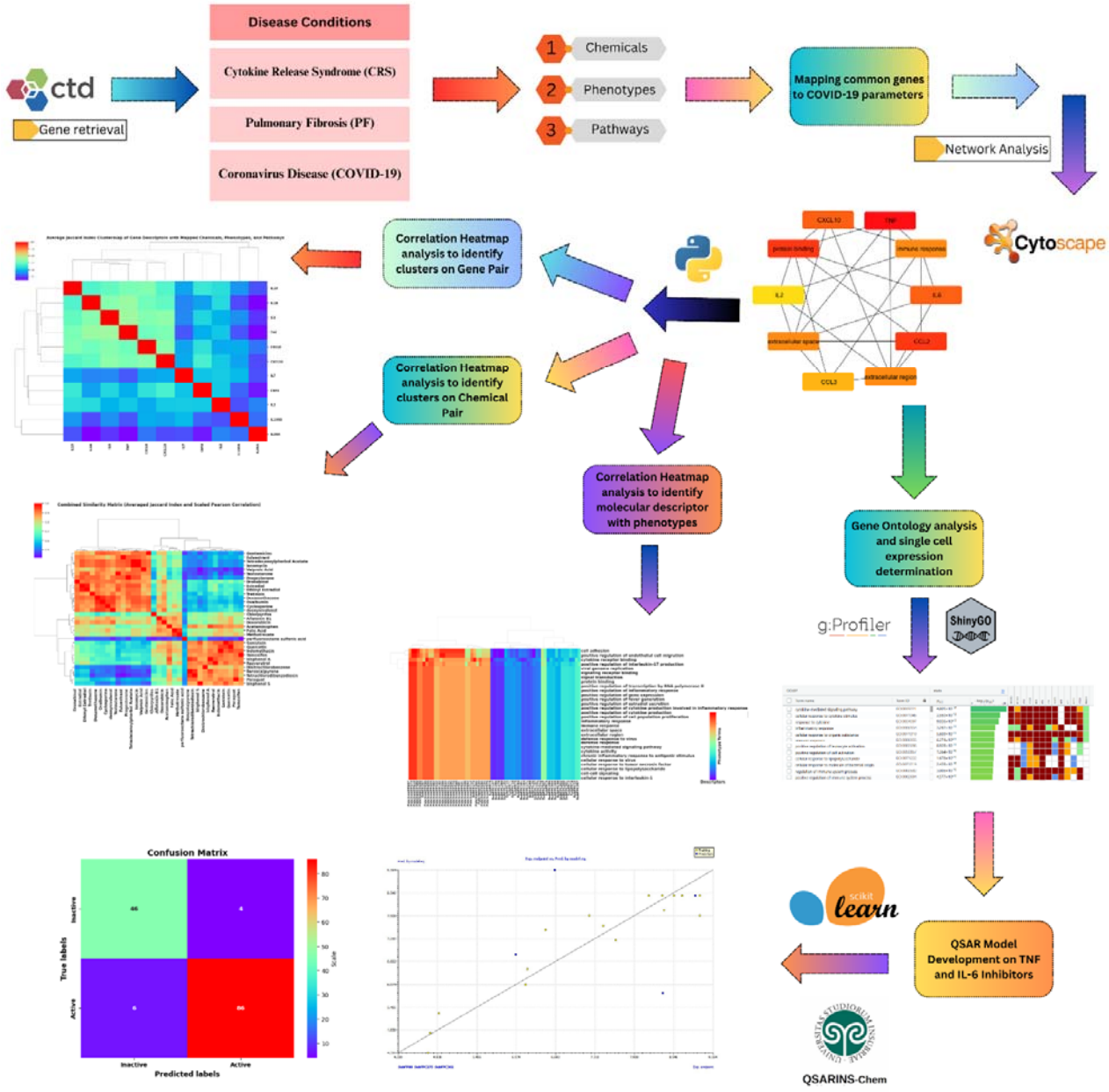
#These authors contributed equally

***Corresponding author**

Ashok Aspatwar

ashok.aspatwar@tuni.fi

Graphical Abstarct



1. Introduction

Coronavirus Disease, also known as COVID-19, has been a cause of intense concern in the medical world due to its novel nature and the absence of any dedicated medical attention or treatments till now. The first case of COVID-19 was observed around December 2019 in Wuhan Province, China. The disease was primarily detected with pneumonia-like symptoms, which were causing high fevers and illness among the affected patients. On further research conducted on the virus, it was immediately identified as a novel coronavirus that belonged to the genus beta coronavirus, which also contained the viruses SARS-CoV and MERS-CoV, which are mainly responsible for the pathogenic development of severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). Upon conducting additional evolutionary analysis on the novel virus in comparison to SARS-CoV, it was observed that there was a sequence similarity of approximately 79% between the two samples. This suggests that the novel virus, which is responsible for a wide range of infections, is biologically related to the pre-existing SARS-CoV virus [1]. Due to its high virulence factor, which was initially unexplainable by medical professionals across the globe, an abnormal rise in COVID-19 cases was observed within a year, which resulted in WHO declaring the infection a global pandemic that lasted until 2022. The progression of COVID-19 involves both severe and asymptomatic conditions, where the asymptomatic progression of the disease involves the development of mild fevers and body pain, which are similar to flu-like symptoms and can be treated with general OTC medications [2]. Conversely, the severe manifestation of COVID-19 infection has been observed to be associated with severe respiratory distress, leading to the emergence of critical infections in the lungs and organ failures that necessitate immediate medical intervention such as mechanical ventilation and intensive care [3][4]. The majority of the mortality rates that were observed among the COVID-19-positive patients were entirely dependent on age factors along with the presence of comorbidities like medical history involving cardiovascular diseases and pulmonary disorders like lung fibrosis, COPD, bronchitis, etc. The majority of cases that required hospitalization of COVID-19 patients were found to be aged individuals with the potential presence of the mentioned comorbidities [5][6].

The pathogenesis of COVID-19 infection primarily deals with the virus interfering with the healthy functioning of the organs and tissues, more specifically localized within the cardiovascular and respiratory systems. The initial infection of the SARS-CoV-2 virus, which causes COVID-19, occurs when healthy individuals inhale the virus in the form of aerosols or droplets from infected individuals. These particles then reach the nasal tissues and infect the upper and lower respiratory tracts of the lungs. The virus actively replicates within the cells of these tissues, initiating the primary progression of the disease [7]. Therefore, the aggressive behavior of the virus causes variations in the overall biological well-being of the affected person due to physiological alterations in the lung tissues caused by the development of COVID-19. This, in turn, results in severe respiratory infections characterized by the production of sputum and excessive mucus discharge from the lungs. These symptoms are assessed through a swab test to confirm the presence of COVID-19 [8][9]. The development of COVID-19 infection in the pulmonary tissues and lungs is

categorized into moderate and severe infection classes based on the level of viral infection in the tissues that affect the normal functioning of the respiratory system [10] [11]. A mild infection is defined by the initial replication of the virus in the cells of the lungs. This type of infection can be treated with over-the-counter medications. On the other hand, the more severe form of infection occurs when the adaptive immune system is activated due to excessive inflammation of tissues caused by the abnormal release of cytokines. This condition, known as Cytokine Release Syndrome (CRS) or cytokine storm, leads to the malfunctioning of lung tissues and has the potential to result in organ failure [12–15]. Cytokine Release Syndrome (CRS) is characterized by the uncontrolled release of cytokines resulting from an infection caused by an overactive immune system and its associated biological pathways. This dysregulation is caused by the negative regulation of genes responsible for cytokine action. Consequently, there is an abnormal escalation of inflammatory responses beyond the normal threshold, leading to cellular damage in the localized infection site. In severe cases, CRS can even result in organ failure [16]. When examining patients with a confirmed COVID-19 infection, it was observed that there was an abnormal elevation in the levels of certain cytokine proteins, specifically IL-1β, IFN-γ, IL-10, and TNF. These proteins play a crucial role in regulating immune system activity. Therefore, it can be concluded that there was a distinct change in the expression of genes responsible for the overall function of these immune proteins [17]. Additional medical research has also noted a rise in the production of IL-6 cytokine levels in individuals who have tested positive for COVID-19, reaching unusually high levels. This increase poses a potential danger of developing acute respiratory distress syndrome (ARDS) and pulmonary fibrosis, which can ultimately result in organ failure [18]. Thus, the advancement of COVID-19 pathogenesis and the occurrence of fibrosis activity caused by CRS may be monitored by examining blood samples and measuring the levels of cytokines to determine the seriousness of the condition. The progression of CRS leading to pulmonary infection is characterized by the hyperactivity of immune cells upon initial exposure to foreign substances or antigens. These cells migrate to various cells and organs, triggering an elevated expression of pro-inflammatory cytokines and chemokines. This leads to the release of abnormally high levels of cytokines due to increased vascular permeability, resulting in heightened inflammation at the site of infection. T cells also play a role in this process, ultimately causing damage to the cells they encounter through excessive inflammation [19].

Pulmonary fibrosis (PF) is a condition where the immune system mistakenly signals for increased inflammation, resulting in the excessive buildup of extracellular matrix (ECM) in the lining of the lungs. This leads to the destruction of the lung's structure and overall biological function [20]. PF is often characterized by age factors, exposure to environmental pollutants, and genetic disorders that affect pulmonary health [21–23]. As previously discussed, the pathogenesis of COVID-19 also includes an increase in the rate of infection among elderly people and through aerosol activity, which potentially contains the virus and is inhaled through the nose, which induces the CRS and hence ARDS. Similarly, we could also conclude that PF could be connected with COVID and CRS due to their likely methods of progression, which could lead to the abnormal release of cytokines within the lung tissues, which could possibly connect to the malfunctioning of the lung ECM and develop into PF [24]. The brief mechanism of

fibrosis is connected with the biological actions that are expressed as a response to the damage to the alveolar epithelium. The alveolar epithelium is the lining of the pulmonary system, which is responsible for the healthy functioning of the lung by facilitating proper gaseous exchange and secretion of surfactant, which maintains lung activity. When these tissues are damaged due to biological stress induced by viral pathogenesis, various inflammatory cytokines are produced on the site of action to promote the healing of the damaged cells, along with the activation of a fibrotic cytokine known as transforming growth factor-β (TGF-β), which plays an important role in the overall PF pathogenesis [25–27]. The TGF-β protein in turn expresses within the infection site and also modulates the expression of αvβ6 integrins, which are another set of proteins that provide positive feedback to facilitate fibrosis activity and restore damaged cells [28,29]. However, as COVID-19 advances and becomes more severe, it can cause changes in biological pathways and the expression of certain traits related to lung processes. This can harm the production of proteins mentioned earlier. Due to abnormal regulation, these proteins can be expressed in excessive amounts, leading to the development of fibrosis conditions. This involves an abnormal increase in cell growth and thickening of the layers inside the cells, potentially resulting in the formation of fibrotic structures within the lung tissues. Ultimately, this leads to pulmonary fibrosis in COVID-19 patients [30]. Therefore, in this manuscript, we present our work by highlighting key concepts related to the biological connections shared among CRS, COVID, and PF, through the implementation of different knowledge graph-based system biology. To identify potential biomarkers, we created a set of hierarchical clustering matrix representations based on knowledge graphs. These representations helped us understand the molecular-level changes in the trisection of the COVID-PF-CRS condition and determine the similarity of gene deregulation by examining the common pathways and altered phenotypes. We also employed a multi-dimensional graph-based method to comprehensively comprehend the mechanism of action of toxicological and therapeutic compounds which involved analyzing their effects on genetic pathways and resulting changes in phenotypes concerning COVID-PF-CRS condition. In addition, we have also created robust quantitative structural activity relation (QSAR) models, along with system biology approaches, to target the most highly expressed hub genes in our knowledge graph. This analysis helped us understand the essential pharmacophores of biologically active small molecules, which can be utilized in the development of drugs for COVID-19 (Table 1).

Table 1: Tabular representation of the computational tools and algorithms used for mechanistic bioinformatics insights

Tools / Methods used	Background of Methodology	Motive of use
Comparative Toxicogenomics Database (CTD)	Cross-reference of data across biological databases and literature	To obtain biological data for respective diseases or biological conditions.
VENNY 2.0	Set theory mathematics	To obtain similar genes among the considered biological conditions.

Correlation analysis for cluster determination	Jaccard Index and similarity matrix.	To obtain a concentration of specific biological properties across gene pairs.
Correlation plots for hierarchical analysis	Neighbor-joining phylogenetic analysis	Performing similarity indexing among molecular descriptors against phenotypes.
Cytoscape/Network X	Graph-based modeling and node-edge interactions	Creation of complex biological networks.
g:Profiler / ShinyGO	Cross-reference of data across biological databases and literature	Perform Gene Ontology and pathway enrichment analysis.
GTEx portal	Cross-reference of data across biological databases and literature	Determining single cell and organ-specific expression among the genes.

2. Materials and Methods

In this manuscript, we utilized a novel bioinformatics pipeline to analyze the spectrum of genes, expressed phenotypes, and biological pathways connecting CRS, COVID-19, and PF. Our methodology involved the identification of the common genes that are involved in CRS, COVID-19, and PF and performing a series of various system biology and graph-based approaches to establish a biological connection to understand the overall correlation of genes and their physiological mechanisms to establish expression in various phenotypes and the pathways they are being involved into. In our study, we used a novel methodology to investigate and study the immunological factors involved in the development of COVID-19 and their relationship to CRS, also known as Cytokine Storm. CRS occurs when cytokines and immune cells are released excessively, resulting in negative effects on the immune system such as inflammation and heightened immune responses. These effects can lead to tissue damage and dysfunction in various organs, particularly in the pulmonary tissues. Our manuscript focuses on studying the impact of CRS on pulmonary infections, bronchitis, and fibrosis. To understand how genes are connected within specific biological pathways and how they contribute to specific phenotypes, we use graph models to visualize these connections. The nodes in the graph represent genes, while the edges represent biological or physiological factors that link these genes together. By examining these connections, we can gain mechanistic insights into the immunological and pharmacological involvement of immune cells in COVID patients, specifically concerning cytokine release syndrome (CRS) and pulmonary fibrosis (PF). This analysis provides evidence to support the hypothesis that negative immune responses in COVID patients are correlated with specific pharmacological factors.

2.1 Obtaining the biological data for COVID-19, CRS, and PF from the CTD database and determining the common genes involved within each condition

The initial objective of our pipeline involved curation of the gene datasets for COVID-19, CRS, and PF respectively from the Comparative Toxicogenomics Database (CTD) [31]. Initially, we curated a set of individual raw genes from each of the specific conditions to form a dataset. The curated dataset was further annotated with the biological pathways and phenotypes associated with the genes concerning COVID-19. Following this we obtained the specific pathways and phenotypes which are involved and expressed respectively within COVID-19 infection which will be used as the main reference to correlate the genetic and physiological factors of both CRS and PF in the network biology process. Using the VENNY 2.0 server [32], we extracted the common sets of genes that are present in CRS, PF, and COVID-19 respectively using the set intersection method. This provided the data points for further biological analysis to establish the physiological significance and significant association to assess the pharmacological reactions associated with COVID-19 pathogenesis following its progression to cause Fibrosis in pulmonary tissues caused by the detrimental effects of Cytokine storm.

2.2 Mapping the common genes against COVID-19 pathways and phenotypes to establish a biological correlation

Following the data curation strategy for obtaining the common genes concerning CRS, COVID-19, and PF respectively into a singular gene set which acted as the entry point for the project, we accessed the COVID-19 disease details from the CTD database and obtained the following classes of data like chemicals, phenotypes and pathways which corresponded to the mechanism and pathogenesis of the mentioned disease. The raw data collected from each specified class included tabulated information for each category, as well as the genes responsible for its function. The chemical dataset comprised a diverse range of chemicals that either exhibit mechanistic action in the progression of COVID-19 or possess therapeutic properties that could potentially help control the disease. Additionally, the dataset included genes that serve as targets for these chemicals in biological processes.

Additionally, the phenotype dataset includes a range of disease symptoms that are expressed as the disease progresses and are crucial for understanding the overall development of COVID-19. This dataset also includes the genes that are responsible for the specific expression of each phenotype. The pathways dataset includes the range of biological pathways and reactions associated with COVID-19 infection, as well as the genes that are closely linked to these biological activities and participate in those reactions.

To link a common gene set list and map it to specific datasets, we employed the Python Pandas data processing library [33][34]. This allowed us to gather information about the chemicals, phenotypes, and pathways associated with COVID-19, CRS, and PF. By doing so, we established a biological connection between these diseases. We then utilized graph-based systems biology methods to create complex

visualizations that illustrate the shared connections among these diseases. Through this analysis, we aimed to understand how their co-expression and shared mechanisms contribute to the immune response in human biology. Additionally, we sought to provide a comprehensive summary of the mechanistic actions and cellular involvement that result from these shared connections.

2.3 Analysing the connections among chemicals, pathways, and phenotypes based on similar gene occurrences and deploying network visualizations

After acquiring the mapped data for common genes related to chemicals, phenotypes, and pathways associated with COVID-19, we proceeded to establish connections between these datasets. This involved linking pairs of data such as chemicals-phenotypes and phenotypes-pathways by comparing the genes involved in each dataset. We determined if there were any similar genes present in the chemicals, phenotypes, and pathways. There is a potential for interconnection between those phenotypes and pathways, which may be related to the mechanism of action of the chemicals as well. Hence by utilizing this logic, we created a custom knowledge graph using graph-based modeling techniques using Cytoscape [35] and NetworkX [36] to analyze the connections among the mentioned classes and establish a biological inference to support the interaction to connect the mechanisms of chemicals and genes which were being responsible for the expression of the phenotypes by involving in various pathways. The nodes in the knowledge graph represent the chemicals, phenotypes, and pathways respectively and the edges connecting them represent the similar genes that were present both along the respective classes represented within the nodes. Thus by this methodology, a well-defined complex network was derived that connected the phenotypes and the pathways concluding the specific phenotypes that were expressed or resulted from the specific biological pathways which in turn would possibly be activated by the expression of certain genes. This pertains to the immunological reactions that occur in COVID patients, leading to cytokine responses. These responses are caused by aberrant concentrations of immune cells, resulting in CRS (cytokine release syndrome) and potentially leading to pulmonary infections and fibrosis if the cytokine storm affects the cells in the pulmonary locations. Henceforth, we constructed several biological networks in the form of knowledge graphs that analyzed the interactions between the genes and the mentioned classes, respectively and obtained the hub networks as well to analyze the interactions that are occurring on a major level. The importance of hub networks involves determining the deregulated genes, mechanistic or therapeutic chemicals, modulated pathways, and phenotype alterations that were mostly concentrated in the network, thus providing biological evidence regarding the specific pathways and phenotypes that are mostly observed at the genetic level in patients suffering from COVID-19 and diagnosed with abnormal cytokine storm and pulmonary fibrosis. Moreover, we interconnected the above networks to observe interactions between respective chemicals and phenotypic alterations to understand the bioactive compounds and substances that are mechanistically connected to facilitate the expression of the respective phenotypes, as well as compounds that could have therapeutic effects involving inhibitory effects to combat the abnormal cytokine storm effects within COVID patients and reduce the detrimental effects on the pulmonary tissues to control the fibrosis of the

cells. We also obtained the gene expression data labels from CTD, where we mapped the genes and chemicals along with the rate of expression levels associated with the chemical action on the gene. Considering both increases and decreases in the expression of genes concerning their simultaneous action by the bioactivity of various chemicals, we also designed a complex biological network that interconnected all the genes, phenotypes, pathways, and chemicals along with the rate of expression as the edge property connecting the genes and chemical nodes, respectively. In addition to the chemical-phenotype network, we also designed the biological interaction network between the pathways and phenotypes to understand the expression of specific phenotypes along various biological pathways that could get activated due to the presence of abnormally high counts of cytokines and other immune cells in COVID patients and possibly lead to the progression of pulmonary fibrosis among the patients.

2.4 Development of a gene frequency matrix for phenotype-pathway associations and creation of gene and chemical pair similarity analysis using the Jaccard Index.

For the next set of analyses, we calculated the frequency count of genes across each chemical, pathway, and phenotype to determine the specific classes with the highest gene involvement. Similarly, we also calculate the presence of the above parameters against respective screened chemical pairs obtained from the CTD to analyze the chemical candidates sharing the majority of the mentioned biological factors that depend on either the mechanistic or therapeutic action of the chemicals. This would enable us to select the phenotypes and pathways that relate to the involvement of most genes along with the mentioned chemicals and compounds that are involved in the pathogenesis of COVID, CRS, and PF. Dividing the following analysis into two parts, firstly we screened out the mentioned classes with a count of genes greater than 10, we performed a similarity matrix calculation across each possible gene pair to analyze the frequency of phenotypes, chemicals, and pathways shared among each gene pair. This would enable us to classify the gene pairs that are most important in terms of the progression of COVID-19, which would lead to abnormal release of cytokines and immune cells leading to CRS and a possible cause for negative responses leading to pulmonary infections and fibrosis of tissues. Further in the following step, we utilized the identical approach to calculate the similarity matrix against possible pairs of the compounds and chemicals we obtained from CTD from which we calculate the occurrences of genes, pathways, and phenotypes to identify the best candidates which are involved in the overall progression and pathogenesis of COVID-19, CRS, and PF being highly correlated with the expressed phenotypes and the biological pathways through differential expression levels of most genes. To create the similarity matrix, we utilized the concept of the Jaccard Index (JI), which is calculated by dividing the common values along the pair of analyzed samples by the total number of values contained by each of the samples in the pair, with the value range lying between 0 and 1. The higher the values are along each mentioned pair, the more concentrations of biological parameters are shared among them. To analyze the results graphically, we developed heatmaps using the matrix generated from the JI matrix to understand the clusters formed based on the values. The deeper the cluster is, the stronger the calculated pair is; hence by analyzing the heatmaps, we could deduce the genes and the

molecular candidates among the best ones that have maximum involvement within the pathogenic reactions that lead to correlation in the progression of COVID and CRS, leading to PF and other pulmonary disorders.

2.5 Development of cluster maps and hierarchical dendrograms to analyze the localization of molecular descriptors of the chemicals along the phenotype classes with the most gene frequencies

Our next analysis involved analysis of the chemicals and bioactive compounds against the expressed phenotypes to assess specific structural and functional properties of the selected bioactive compounds and substances that could be involved in a mechanistic manner, which would mean that the respective compounds would be working in a correlated manner to induce the pathogenesis of COVID-19, initiating the CRS and leading to damage in pulmonary tissues leading to PF. On the other hand, the bioactive compounds could also demonstrate a therapeutic nature, which would have inhibitory effects against the pathogenesis of COVID-19 and hence could be used to create drug discovery approaches to control the abnormal cytokine release that leads to medical emergencies like fibrosis in organs, more specifically the lungs, as per our project. Thus, deciphering the molecular mechanism of chemicals obtained from the CTD after mapping the genes along COVID as a reference is extremely important to screen out the chemical candidates that showed a higher interaction potential against the expressed phenotypes. To make the study more specific in terms of chemical morphology, we deployed the utilization of the concept of molecular descriptors and fingerprints, which are the physiological characteristics and maps of chemicals and compounds, which provide us with overall information regarding each specific structural morphology class of the specific compound. Hence, while assessing each chemical against its expressed phenotype, we could analyze exactly which molecular fingerprint and descriptor would correspond significantly against the phenotype, which would give us a detailed explanation regarding the chemical bonds or functional groups specifically responding to that particular fingerprint, hence providing us with a more localized approach towards screening the molecules. It’s a known fact that specific functional groups and atomic constituents within a molecule or compound can exert significant biological reactions on the parent compound. Hence, this property is used as a method to perform a functional group analysis in the drug discovery method, which involves the observation of correlations between these specific fingerprints and the net biological activity against a phenotype or disease. After extracting the PubChem and Substructure fingerprints from the PADEL descriptor [37] package for each screened compound, we mapped them against the respective phenotypes they corresponded to and developed a clustering heatmap combined with hierarchical dendrograms to observe the similarity relationship among each descriptor and how they related to the specific phenotypes. The cluster heatmap and dendrograms combined not only provided us with the mean concentration of significant molecular fingerprints, which are associated with the biological expression of the phenotypes but also the biological similarity and mathematical correlation values between the fingerprints and the phenotypes, which would prove to be an informative approach to conducting a further study about the chemicals and their respective fingerprints and how they would relate mechanistically in the progression of the phenotype class. The clusters obtained within the heatmap suggested the similarity and the correlation among the molecular

fingerprints against the specific phenotype, which would also be strong evidence to suggest its involvement in expressing the phenotype along the pathogenesis and progression of PF due to cytokine storm in COVID patients.

2.6 Analysis of best-involved gene sequences from the pathway enrichment analysis and obtaining the best chemical candidates through JI heatmap calculation.

From the procedures involving graph-based system biology approaches mentioned previously, we obtained the set of genes and chemical candidates from the JI heatmap and the clustering from pathways, phenotypes, chemicals, and morphological descriptors, they are mechanistically connected with. Similarly, the JI correlation matrix across the chemical pairs was performed to deduce the occurrences of major genes, pathways, and phenotypes that would possibly be involved within the mechanistic mode of action of the respective chemicals and compounds when exposed to the biological system. The analysis highlighted that the genes and the chemical candidates obtained from the JI heatmap have been shown to occur more frequently than their other involved counterparts in the network. Therefore, this new set of genes and chemicals we obtained gave us a newer and more detailed insight regarding the specific types that are mostly involved through their mechanistic activity within biological interactions and pathways that are involved in the overall pathogenesis of COVID-19 as well as contribute towards the CRS symptoms, which in turn could lead to the development of PF conditions in humans. To support our hypothesis and verify the results we obtained from the graphical methodologies as explained in the earlier sections, we analyzed the gene sequences using gProfiler [38]and ShinyGO [39]to obtain detailed information regarding the Gene Ontology (GO) terminologies, pathway enrichment information, along with cellular localizations, including the chromosomes the set of genes belonged to. These series of analyses and the obtained results further verified the inferences we had previously produced from our graph-based system biology methods. Additionally, we also extracted the biological pathway responding to COVID-19 infection and labeled the extracted genes to understand their involvement in the overall pathogenesis of the disease. The chromosomal localization map for the gene set also provided an informative insight into the gene position within the human genome system and hence could be a potential factor within transcriptomic analysis where the specific chromosomal number is required to map the biological samples to analyze the differential gene expression analysis. Moreover, we also determined the organ-specific localizations for each gene using the GTEx portal [40] to understand which organs specifically the screened genes from the entire analysis are being expressed. Additionally from the JI similarity matrix we obtained by clustering the genes, phenotypes, and pathways across the chemical and molecular compound candidate pairs, we obtained specific clusters from the heatmap based on the higher range of correlations denoting there was possibly a connection concerning the mechanism of action which is shared by all the considered chemicals that were observed within each separate clusters. Therefore to obtain further information regarding those clustered chemicals and compounds and understand the biological reason behind their correlation within the cluster, we referred to drug databases like PubChem and ChEMBL to visualize their mechanism of action within the biological

pathways we are targeting along with their overall physiological information which would enable us to understand the mechanistic behaviour of those chemicals and how they affect the overall pathogenesis of COVID and contribute to the progression of PF through CRS activation. This would provide us with a clear idea not only about gene expression and pathways shared by the genes but also about the organs and tissues where they are differentially expressed along with the detailed mechanism of action of the above-mentioned chemicals and molecular candidates and their involvement to facilitate the expression of the top performing genes within the biological pathways ultimately controlling the expression of various phenotypes associated with the development of PF within COVID patients. Hence, the results obtained from this would further support the hypothesis of the progression of CRS into pulmonary tissues, which is the main aim of our manuscript.

2.7 **Development of Quantitative Structure-Activity Relationship (QSAR) models for anti-TNF and anti-IL-6 compound candidates from the ChEMBL Database**

From the analytical approaches we discussed previously, we obtained a wide spectrum of data supporting the correlation of top-expressing genes and various chemical candidates and compounds that play an integral role in the overall pathogenesis of COVID-19 and contribute to the progression of PF through CRS activity. By utilizing data mining techniques in various available research literature and consulting drug and gene databases, we were able to establish a significant connection between highly correlated genes and chemicals, as well as their mechanistic mode of involvement. This connection facilitated the initiation of a set of biological pathways that either contribute to the disease's progression or provide therapeutic actions to prevent the progression of the disease's associated symptoms. In the following analysis, we leveraged the ability of machine learning (ML) approaches to developing QSAR models using the Sci-kit learn package [41]and QSARIN software [42]to perform in silico screening of bioactive compounds and potential drugs from the mentioned chemical databases, which are known to possess potential inhibitory activity against the TNF and IL6 target genes within the immune system. Within COVID patients, the progression of CRS is often accompanied by the abnormal secretion of mostly TNF and IL6, which contributes to the deployment of cytokines at the site of infection, where inflammatory responses are initiated by immune reactions, which at an abnormal level lead to the formation of fibrotic structures into the pulmonary tissues and the alveolar ducts, hence leading to the PF. Therefore, the methodology to screen out potential inhibitors to block the abnormal activity of TNF and IL6 and prevent the progression of PF phenotypes in COVID patients is actively required. Based on the availability of data for each of the targets, both classification and regression QSAR models were developed, which classified the active chemical candidates and the inactive chemicals based on their range of bioactivity (pIC50) values and predicted the pIC50 values for a set of compounds, respectively. To develop the model, the features used as markers for each chemical candidate were the molecular descriptors from the PADEL toolkit, which included the PubChem fingerprints and the Substructure fingerprints, respectively, which are the various morphological and physiological properties of the chemicals, which are denoted in a binary format based on their presence or absence. The potency of

chemicals is often dependent on their morphological properties, which are individually responsible for exerting certain biochemical reactions within the biological pathways, which can either play a mechanistic mode of action to promote the progression of the disorder or provide a therapeutic response to relieve the stress induced by the abnormal expression of target genes or proteins, which are responsible for the overall pathogenesis. Thus, we establish a statistical relationship utilizing the descriptor properties to predict the biological activity in the form of inhibitory concentration values to screen out potent chemical candidates and deduce the potential set of compounds that can be used to tackle the abnormal expression effects caused by TNF and IL6 within the immune system. The algorithms we concentrated our main focus on were Random Forest (RF) and Multiple Linear Regression through Genetic Algorithms (GA-MLR), which are known for their robustness and ability to process complex datasets without the risk of potential data loss. RF and GA-MLR utilize complex feature selection workflows to work on smaller datasets to facilitate statistical connections by creating decision tree structures and imitating genetic selection techniques, respectively, which are responsible for finding detailed connections with every individual chemical feature to deduce the most important features that were responsible for the overall prediction of biological activity and were statistically significant in the creation of the entire model. To visualize the accuracy of the models, we plotted the model performance in the form of a confusion matrix, ROC plot, and Important Features plot to understand the positive performance and the classification ability to differentiate between potent and inactive chemical candidates. Additionally, for the regression model, we utilized the concepts of applicability domain and linear relationship plots to verify the predictive capability of the GA-MLR algorithm and also obtain the best-performing chemical candidates to further process them to analyze the reactions between their target genes and the respective pathways and phenotypes associated with the pathogenesis of COVID-19 leading to PF. Ultimately, after screening out the best-performing chemical candidates and extracting their important descriptors, which were found to be significant in the overall prediction process, we extracted the target genes from the Swiss Predict server, after which we mapped the gene with the highest probability factor to our created dataset containing the expressed phenotypes and the biological pathways associated with the overall pathogenesis of COVID-19, which leads to PF due to CRS development. Hence, from the following analysis, we successfully constructed a predictive model to classify and predict the active compounds along with their bioactivity value, which can act as potential inhibitors to provide mechanistic action on TNF and IL6 to control their abnormal expression effects and protect the immunological health of the patient.

3. Results and Discussion

We generated the results by analyzing as mentioned in the previous section and got detailed results depicting the biological connection shared among COVID-19, CRS, and PF, along with the set of genes that were actively involved in the differential expression of various phenotypes, along with the pathways and biological reactions they were involved in. Additionally, we also analyzed various sets of chemicals that were exhibiting mechanistic action in the biological pathways and promoting the disease progression or

showcasing therapeutic action that would contain the potential to inhibit the pathogenesis of COVID-19 and assist in the control of the release of cytokines, reducing the risk of tissue degeneration and fibrosis. In the subsequent sections, we will provide a comprehensive analysis of the interactions between genes, pathways, chemicals, and phenotypes within various networks. Specifically, we will focus on the significance of these interactions in establishing biological connections and promoting the mutual development of COVID-19, CRS, and PF pathogenesis in humans. In addition, we conducted a thorough analysis to examine the relationship between the chemicals mentioned and the observed phenotypes. We utilized hierarchical clustering and a correlation matrix to identify the specific molecular fingerprints that are associated with each phenotype. This allowed us to identify the molecular features or scaffolds that may be responsible for the development of disease symptoms. These findings could be valuable in designing drug candidates and developing methodologies for treating patients with abnormal cytokine levels in COVID-19, to prevent medical emergencies such as organ failure and pulmonary disorders, which are commonly targeted during CRS progression.

3.1 Collection of common genes from COVID-19, CRS and PF

As discussed in Section 2.1, we collected the genes responding to the mentioned biological conditions separately and extracted the common genes that corresponded to the pathogenesis of COVID, CRS, and PF. This set of common genes would be the key values that we would be using to perform connections against the pathways and phenotypes of our reference disease, i.e., COVID-19, and which in turn would be used to construct complex biological networks and cluster maps to analyze the chemical action in a mechanistic or therapeutic manner. Figure 1 below provides the Venn diagram representation of the common genes obtained by intersecting the separate gene sets from the three conditions, respectively.

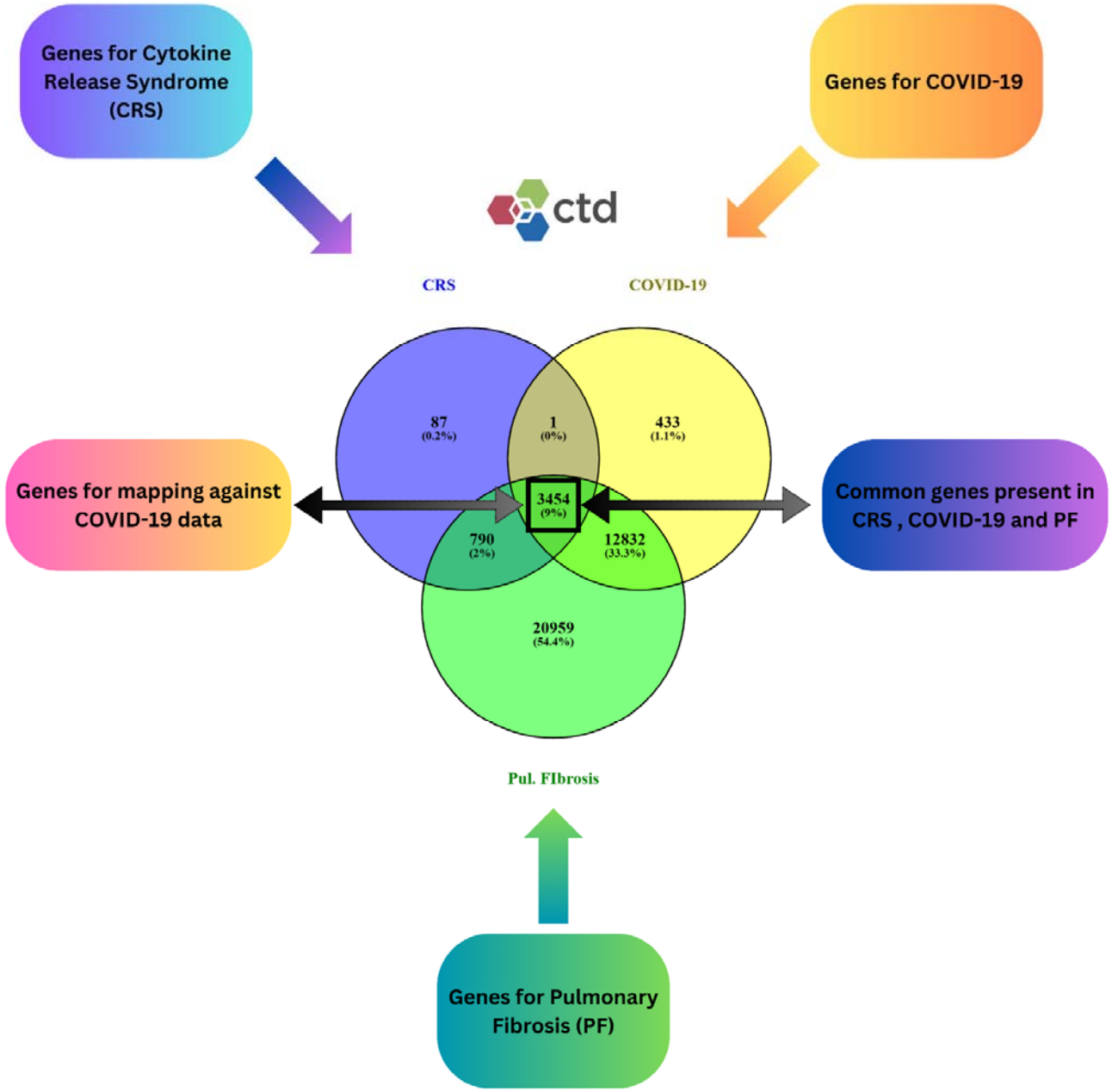


Fig. 1: Visualization of the common gene set among COVID-19, CRS, and PF

3.2 Visualization of biological networks to the interactions between common genes and the COVID-19 parameters.

We had mentioned in Section 2.2 that after obtaining the common genes that were shared among COVID-19, CRS, and PF, we mapped them against the biological data of COVID-19, which contained the data for pathways involved in its pathogenesis, the pathways that were expressed during the progression and development, and the chemicals that were involved in either mechanistic action and acted as a promoter to facilitate the pathway or were responsible for carrying out inhibitory effects against the virulency of the disease. The main motive of this step is to establish a complex connection between the genes that are

responsible for cytokine storm progression due to the COVID-19 infection, which causes the abnormal release of cytokines and other immune cells, which leads to the initiation of immune responses, causing the inflammatory syndrome, which, when occurring at an abnormal rate, can develop into damage to tissues and organs, affecting the cellular areas with a higher amount of cytokine concentration. As our manuscript is focused more on the progression of fibrosis in tissues of the lung and other pulmonary cells, we aimed to connect the gene data along the pathways to understand the phenotypes that were expressed in patients with pulmonary fibrosis disorder who had a COVID-19 infection. Therefore, we constructed the connections using interactive biological networks in the form of knowledge graphs, with the nodes representing the genes involved in all the mentioned diseases and conditions and the edges representing the connection or interaction factor, to understand which genes correspond to which particular pathway and similarly, which specific phenotype classes were expressed and observed significantly among patients with lung fibrosis with a positive diagnosis of COVID-19. Additionally, we also constructed the network to analyze the chemicals obtained from the COVID-19 dataset and how they interact with the gene set to get a descriptive idea about the chemicals or substances that influence the actions and overall expression of the genes associated with CRS and PF pathogenesis. The networks were constructed using Cytoscape and NetworkX, and hub network identification was performed based on the degree score method and closeness scores to obtain the best components, which were correlated as well as present at majority levels within the overall biological progression of the mentioned conditions. The networks constructed among the genes with their respective chemical, phenotype, and pathway counterparts are shown below respectively in (Figures 2-5) (Supplementary S1-S3). Additionally, the full network representations for the entire biological networks are provided as well along the section.

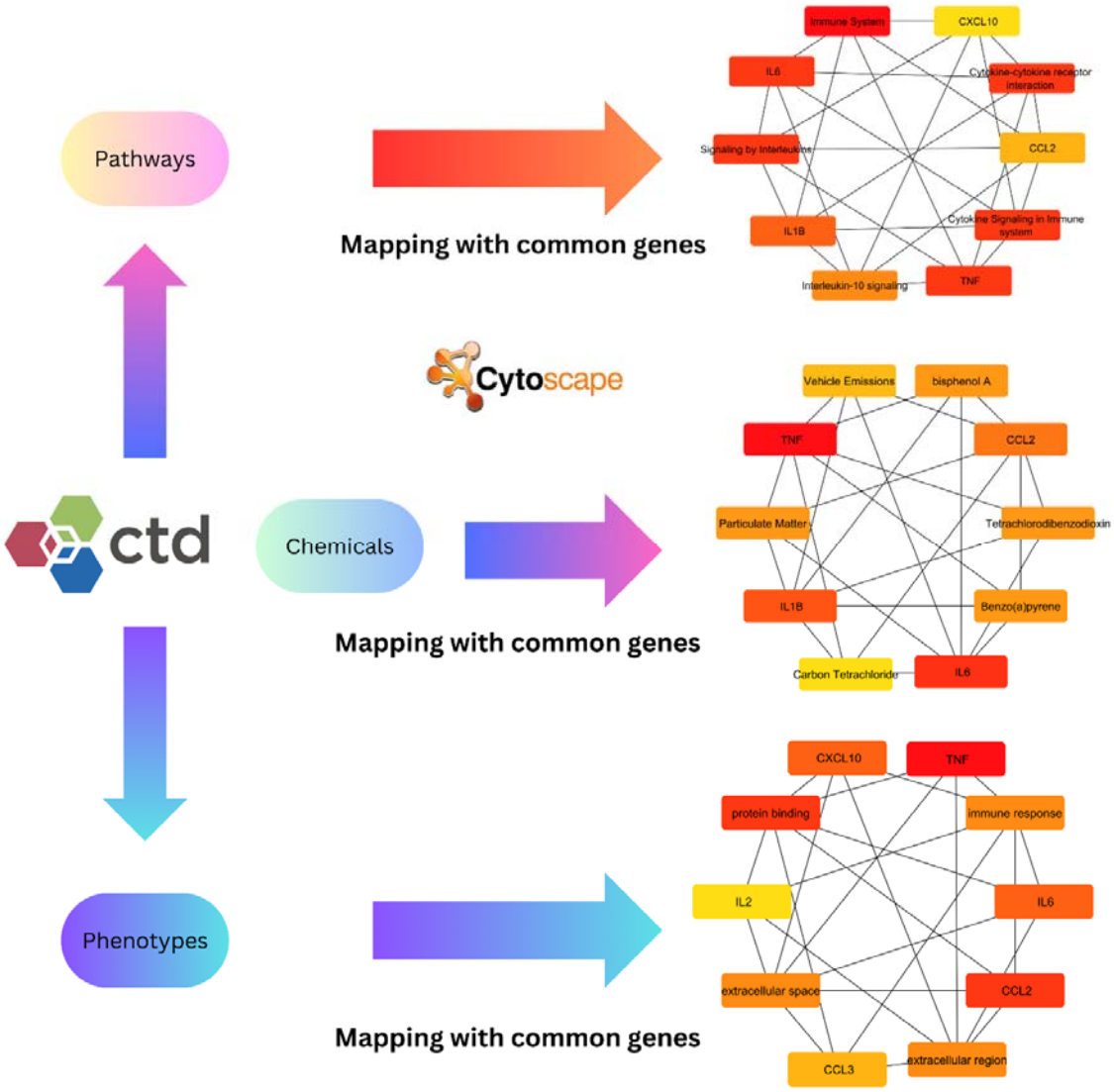


Fig.2: Visualising the hub network between common gene and expressed phenotypes from COVID-19

From the given figure (Fig. 2), we observed specific biological interactions among the genes TNF, IL2, and IL6, which are more commonly known as tumor necrosis factor and interleukins, respectively, which belong to the class of proteins or genes involved highly within immunological responses and play an important role in facilitating immune reactions during infections or viral fevers. These genes facilitate the release of cytokine proteins, which act as signaling molecules to appoint other immune cells to assert repair functionality and inflammatory responses at the site of infection and viral attacks, hence contributing to treatment processes. Moreover, it can be observed that among the phenotypes expressed, immune response and protein binding are highly correlated with the mentioned genes, from which we could conclude that there is indeed a biological connection among the progression of CRS among COVID-19 patients, which occurs due to abnormal immune responses and inflammation along the pulmonary tissues beyond healthy

levels, ultimately leading to fibrosis along the cells. Additionally, the extracellular space and binding that were also observed within the diagram could depict the cellular regions along the borders where the COVID spike protein would bind to initiate the immune response and hence progress forward with the cytokine action, which could, due to biological anomalies, result in the above emergencies as discussed. Moreover, we observed the interactions among various chemicals and substances with immune genes like TNF and IL, from which we could infer their mechanistic relations along the progression of PF within COVID-19 patients. From the network initially, we saw the connections drawn among vehicle emissions, particulate matter, and immune genes like TNF, IL, and CCL. This relationship could be due to the substances being classified as air pollutants, or more specifically, allergens, which are responsible for activation of the immune system, which in turn could lead to the expression of genes like TNF and IL, which are the integral components for initiating the cytokine response in the form of hypersensitivity reactions and inflammatory responses, which in severe cases could take the form of CRS. During the peak of COVID, it was also a concern for patients with excessive allergic reactions, as certain of them infected with the virus also showed symptoms related to hypersensitive reactions at severe levels, which were characterized by abnormal immune cell concentrations in the system. Hence, there's also a possibility that these inflammatory responses on a severe level might lead to an increase in cytokines and related immune cells along the infection sites, which could lead to PF conditions as the air pollutants reacting with the immune system are localized along the lung tissues only as they are inhaled. Moreover, certain chemical samples like benzo(a)pyrene, tetrachlorodibenzodioxin, and bisphenol A were also observed within the network, interconnecting with the respective genes, which could be potential chemical candidates that would have a possibility of being involved in a mechanistic manner in the progression of COVID-19 and connected to facilitate or act on the target proteins or genes that are specifically responsible for CRS and PF. Hence, from this network, we obtained certain substances and biological compounds that we could use as potential samples to analyze the biological response activity of the infection caused by COVID-19 and CRS. Additionally, the representation also derives the interactions among the common genes and the biological pathways associated with COVID-19. Based on the explanations we provided for the previous diagrams involving the phenotypes and chemicals, the pathway interaction is similar in terms of analysis. In this result, we observed mainly the immune system and the cytokine-mediated pathways to be connected with the immune genes along with the interleukin pathway, which altogether concludes the fact that the genes we analyzed from CRS, COVID, and PF samples are significantly influenced and actively participating within the activation of the immune system and more specifically account for the cytokine signaling action, which is solely responsible for the progression of CRS and hence leads to the development of lung infection due to excess cytokine activity and causing PF. Additionally, the interleukin pathways that can also be observed in the network verify the fact that the deployment of IL proteins is also occurring, which work biologically alongside TNF to initiate the action of cytokine release along the region of infection and develop the inflammatory response to counter the immune reactions.

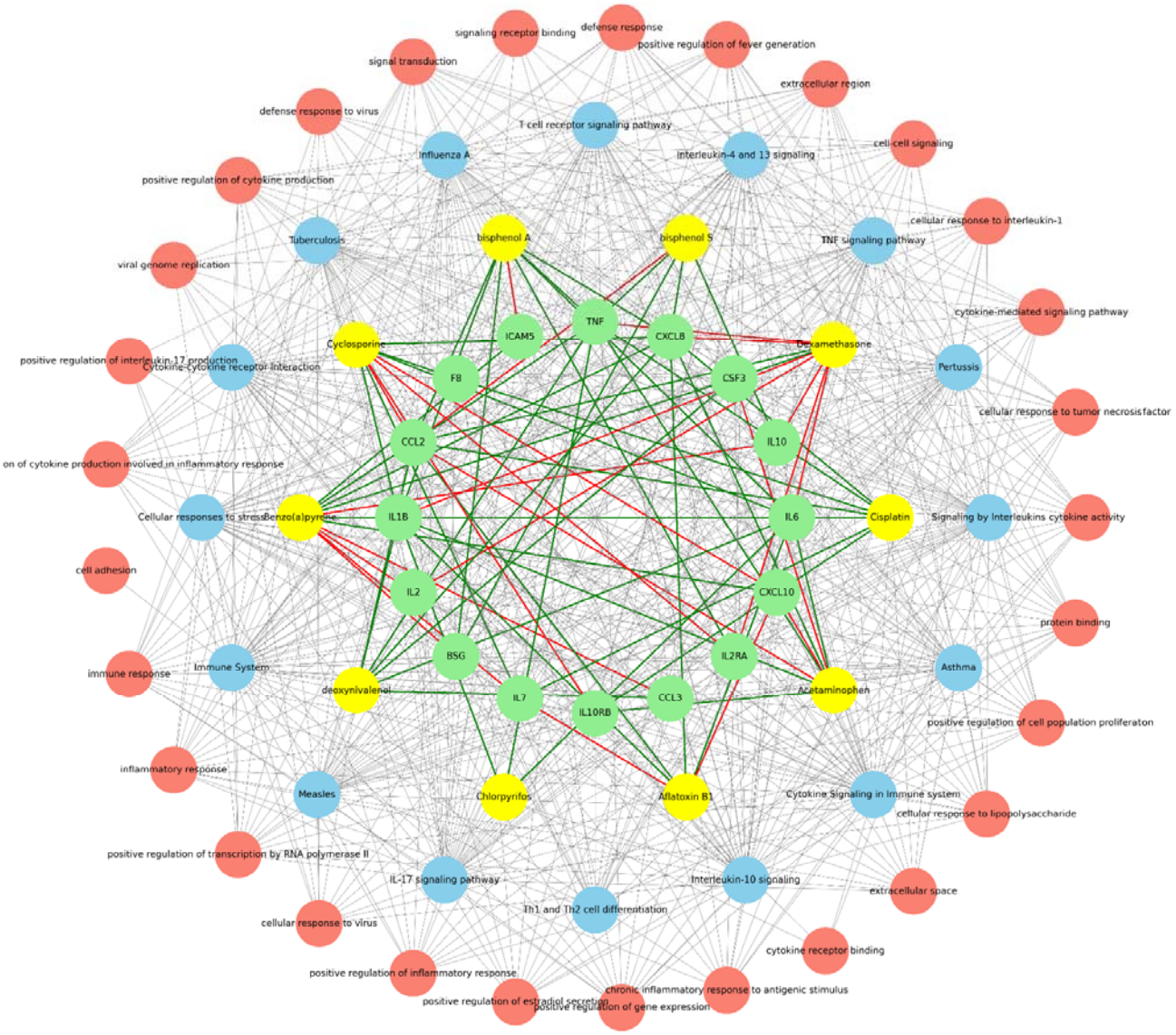


Fig.3: Visual representation of the knowledge among the deregulated genes, expressed phenotypes, and biological pathways associated with COVID-19

From the above representation (Fig.3), we observed the complex biological interactions established among the common genes obtained from COVID-19, CRS, and PF, respectively, along with the expressed phenotypes, biological pathways, and specific chemical components associated with COVID-19 pathogenesis. The genes are represented as green nodes, the biological pathways as blue nodes, the phenotypes expressed by their correlation as red nodes, and the chemicals as yellow nodes. From the network, we could conclude that the interactions formed among the mentioned components are highly interconnected with each other, which implies that almost all the biological activities, along with the expression within the biological pathways, are connected for the combined activity of the genes, which verifies the mechanistic action of all the genes being an integral factor in the development of PF progression

established among the common genes and their respective phenotypes expressed throughout their progression, along with the chemicals and substances that are mechanistically connected with the overall biological processes associated with the expression of the corresponding phenotypes, Therefore, another important approach that we implied from this network was to perform further analysis by considering a set of certain compounds that were found to be interacting with the majority of the phenotypes, utilizing the extraction of various molecular descriptors, which are numerical representations of the morphology and physiology of the compounds, and creating a correlation matrix against the phenotypes to analyze the most important descriptors, which were mostly involved with the biological expression activity of the phenotype. Additionally, we also established the interaction between the chemicals and the corresponding genes concerning their gene expression levels in their biological interaction with the respective chemicals with green edges denoting an increased expression and vice versa for red edges. Since various chemicals and biological substances, due to their physiological properties, exert certain biological activities that overall control the expression of genes within the biological pathways, the representation provided by our network would enable us to get detailed information regarding not only the complex relationships among genes and their respective factors but also their rate of expression when exposed within the biological environment. Consequently, this approach enabled us to screen out chemical candidates that could act as potential drugs for providing therapeutic action against the development of disease phenotypes, thus delving into the concept of drug design methodology to treat PF progression among COVID patients.

3.3 Determination of frequency of gene counts across the COVID-19 biological data and construction of JI heatmaps to perform gene and chemical pair analysis.

As per the analysis we had discussed previously in Section 2.4, after mapping the common genes to the COVID-19 dataset containing the details regarding the chemicals, phenotypes, and pathways respectively, we performed a frequency count of the genes involved in each of 3 mentioned categories of the dataset to determine the specific phenotypes and pathways which are enriched or connected by the most number of genes. This count data enabled us to narrow down our search about the biological functioning and processes and obtain the specific biological pathways and the phenotypes that were selectively expressed within patients affected with COVID-19 and as a possible risk of developing PF due to the abnormal release of cytokines and inflammatory responses due to CRS progression. The following bar plot representations (Supplementary S11 Figure 1s-2s) (Supplementary S4-S5) given below are the respective biological pathways and expressed phenotypes that were observed within the COVID-19 data corresponding to the frequency of shared genes along with CRS and PF respectively. On the other hand, we also constructed the JI similarity matrix among the chemical pairs to calculate the clustering of the respective genes, pathways, and phenotypes when they are mechanistically or therapeutically connected. Moreover, we also calculated the molecular descriptors in the form of PubChem and Substructure fingerprints which represent certain morphological and physiological structures that are present within each chemical candidate. Hence performing a JI clustering method to obtain similarity matrix maps would indicate the similarity of a set of

certain candidates that share similar morphological characteristics and hence would contain a similar mechanism of action against their involvement within correlated biological pathways and phenotype expression which correspond to COVID-19 progression leading to PF through CRS development. Furthermore, we analyzed the clusters formed to deduce the common biological property and its effect within the human immune system whether they provide a mechanistic mode of action to increase the progression of the disease or are responsible for inducing therapeutic response by reducing the detrimental impact of abnormally expressed genes and malfunctioning biological pathways. We used data mining methodologies like literature reference and drug database information retrieval systems to understand the mechanism of each chemical and compound in each cluster and connect their effects to the overall progression of PF within COVID-19 patients.

From the bar plot frequency representations (Supplementary S11 Figure 1s-2s), we observed the specific pathways and phenotypes that had a higher frequency of genes than the others. This implied that pathways like the immune system, cytokine signaling, cytokine receptor binding action, interleukin pathways, etc. are much more likely to participate during the progression of cytokine storm and fibrosis in lung tissues, which was biologically also significant since immunological research explains the cause of CRS to be abnormally activated immune system and receptor binding action only, which cause false signaling of cytokines to the site of infection, which leads to excess inflammatory responses. Moreover, the phenotype representation provided us with the information that an active expression of biological phenotypes like protein binding, extracellular activity, cytokine action, cytokine-mediated signaling, etc. occurred more frequently than others, which implied that the patients suffering from PF due to CRS were diagnosed with the mentioned symptoms or biomarkers when tested positive with COVID-19. Therefore, the following analysis provided us with a detailed explanation regarding the biological activity and specific phenotypes that could be expected in patients diagnosed with COVID-19 and having possible risk factors for PF due to abnormal cytokine activity.

Additionally, we calculated the JI similarity scores across the experimented gene and chemical pairs obtained from CTD to analyze the concentration of biological pathways shared between them along with the expressed phenotypes involved within their biological processes. The main aim of this step was to screen out the set of genes that were more expressed than the others within the biological pathways of COVID-19 along with their interaction with various chemicals within the database that were mechanistically or therapeutically associated with the CRS progression and development of PF. Additionally, we performed a similar set of similarity analyses across chemical pairs where the clusters we obtained provided us with detailed information regarding the various biological pathways and phenotypes shared among them for their mechanism of action and also the target genes they significantly interact with to influence the respective biological reactions. Hence, by utilizing clustering methods, we performed groupings of all the major pathways, genes, chemicals, and phenotypes we obtained in the previous step concerning their corresponding pair analysis for the mentioned factors. For interactive visualization of the JI cluster results,

we developed heatmaps using Python to observe and infer the most correlated gene and chemical pairs with the greatest number of pathways and phenotypes along with the top-performing biological pathways and expressed phenotypes within the clusters with the highest JI scores, which are provided in the following figures (Fig. 4-7) (Supplementary S6-S8).

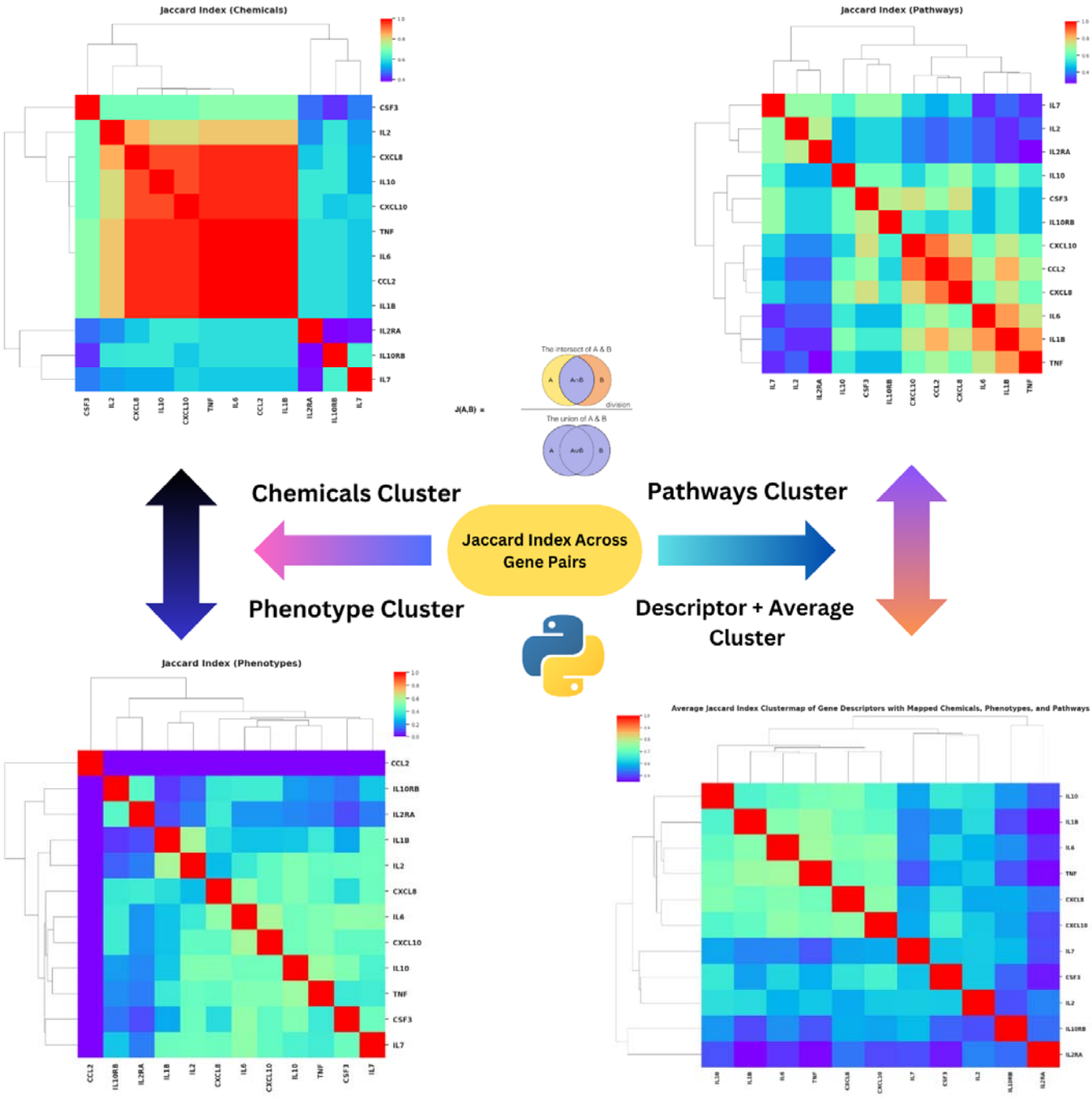


Fig.4: Jaccard Index (JI) of COVID phenotypes, pathways, and chemicals and averaged with molecular descriptors across gene pairs

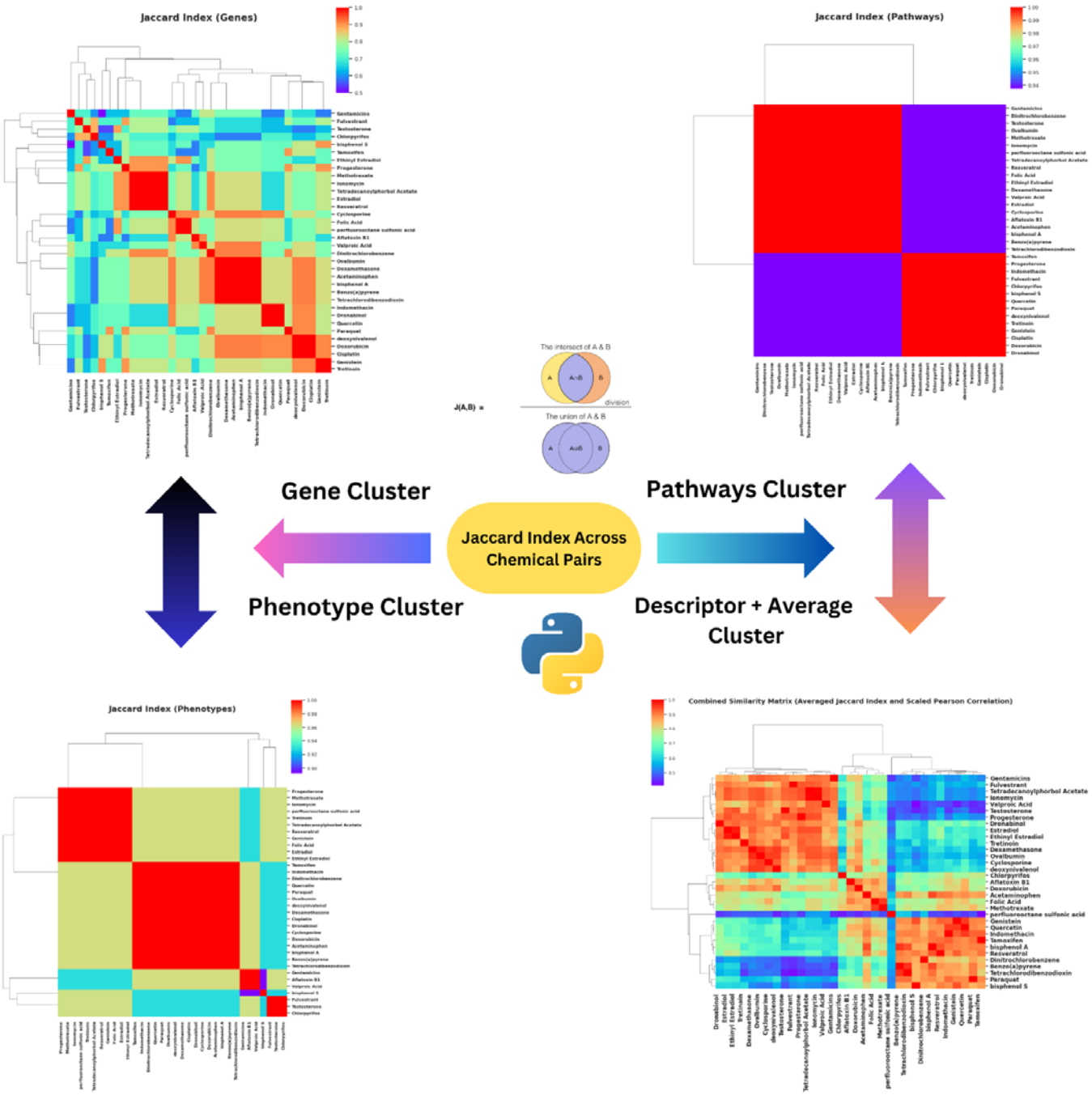


Fig.5: Jaccard Index (JI) of COVID phenotypes, pathways, and genes and averaged with molecular descriptors across chemical pairs

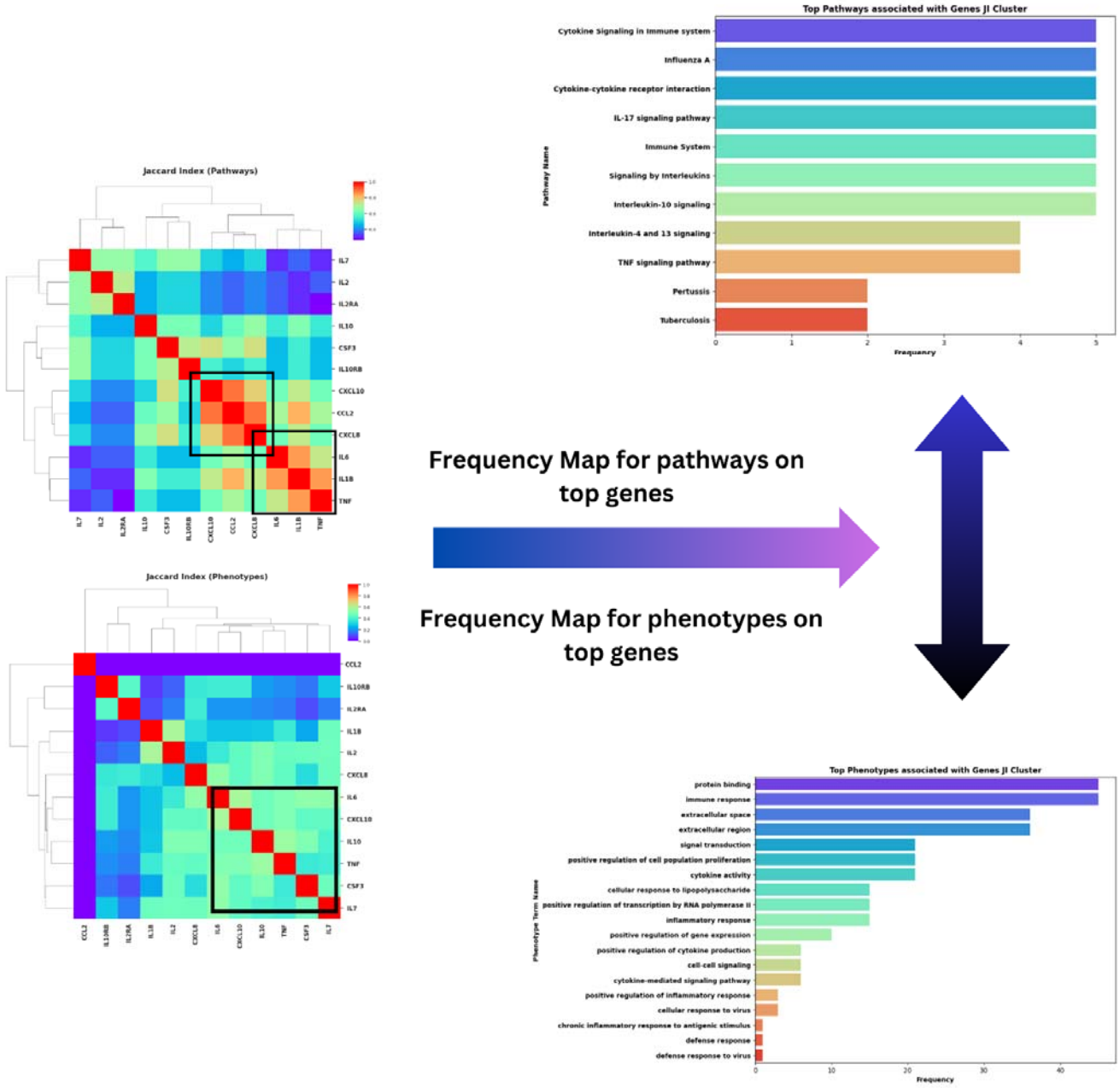


Fig.6: Cluster Analysis to extract the respective pathways and phenotypes corresponding to the gene pairs with the highest JI correlation.

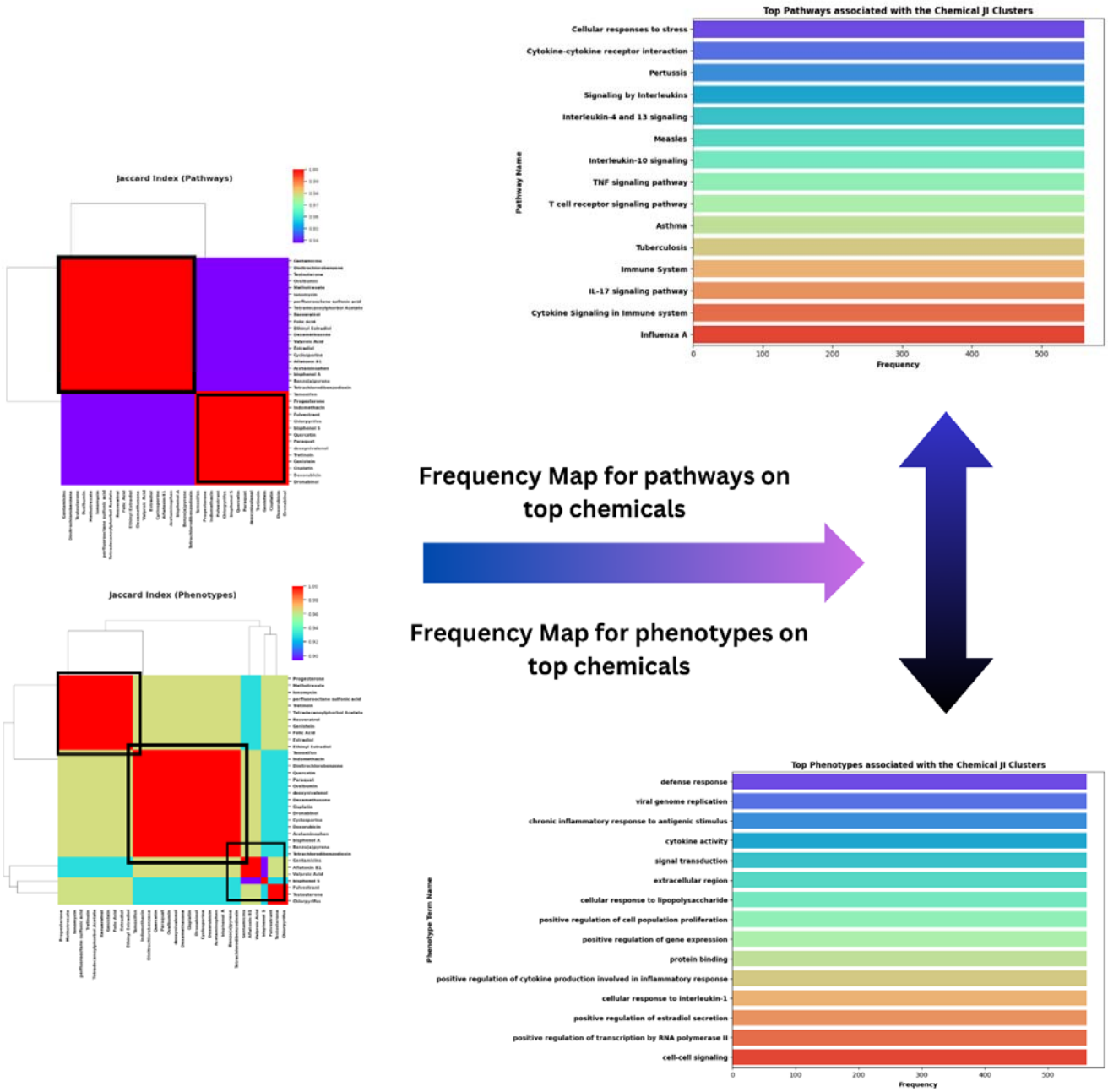


Fig.7: Cluster Analysis to extract the respective pathways and phenotypes corresponding to the chemical pairs with the highest JI correlation

From the above diagrams (Fig 4-7), we observed a high correlation cluster formed among the following genes: CXCL10, CXCL8, TNF, CCL2, IL1B, IL6, IL2, IL2RA, IL10, IL7, CSF3, CCL3, and IL10RB, which denoted that the majority of the pathways utilize the following genes, which would conclude the fact that there was indeed a differential expression of the mentioned genes within COVID-19 patients with risk factors associated with the development of PF due to CRS. Moreover, to support this result, we also performed gene ontology (GO) in the following section using the obtained genes to understand their molecular functions and cellular localizations, along with the organs they are mainly associated with during the development of the disease and fibrosis situations. Additionally, within the JI heatmaps of phenotypes, we observed a specific correlation pattern similar to pathways, which implied that a similar set of genes was potentially responsible for the expression of the phenotypes by their involvement within the respective biological pathways. On the other hand, we observed a lot fewer clustering patterns among gene pairs within the chemical JI map. On analyzing the chemicals specifically along the clusters marked with red color showing the highest JI score, we observe a class of toxins and chemicals that are potential environmental pollutants like Benzo-a-pyrene, Tetrachlorodibenzodioxin, Dinitrochlorobenzene, BisphenolA, Resveratrol, etc, which would possibly provide concrete evidence that they share a similar mechanism of action through their morphological similarity and target the similar biological pathways and genes which would then be responsible for initiating the overall progression of CRS within COVID patients which eventually would lead to PF due to tissue damages as negative feedback from the mechanism of action of those chemicals. On the other hand, we also observed cluster formation among certain chemical candidates like doxorubicin, acetaminophen, Doxycycline, etc., which are known for their therapeutic mechanisms of action against inflammatory responses and a series of infections localized around respiratory tissues and other organs. Hence, performing a cross-checking analysis of the above chemicals from PubChem and associated literature, we depicted the scientific reasoning behind their cluster formation and concluded their significant mechanistic methodology to promote or treat the disease conditions targeting the mentioned pathways. However, the set of gene pairs that showed signs of higher correlation within the JI map consisted of a similar set of genes we had observed in the pathways and the phenotypes as well. Additionally, we also included the functional and morphological descriptors of genes and chemicals respectively due to their significant involvement within the overall biological functions associated with pathways associated with the mentioned diseases. Hence, after combining the three correlation plots after analyzing the clusters and observing the gene pairs that showed maximum similarity scores, we could without any doubt connect the information about the joint mechanistic approach of the chemicals and their involvement within the biological pathways to express the phenotype conditions and also successfully establish the biological conclusion that the chemicals might also act as potential biomarkers that influence the initiation of the biological pathways associated with COVID-19 pathogenesis, leading to the development of PF through CRS progression.

3.4 Development of hierarchical cluster dendrograms to analyze the correlation between chemicals and phenotypes using molecular fingerprints.

As we mentioned in Section 2.5, we performed a detailed analysis to determine the morphological parameters of the various chemicals and their correlation against the expressed phenotypes to derive conclusions regarding their mechanistic or therapeutic properties, which could act as a potential starting point in drug discovery processes. After obtaining the list of chemicals by mapping the common genes from CRS, COVID, and PF to the dataset of chemicals corresponding to COVID-19, we selected certain chemical candidates that are known across the literature and research as potential inhibitors of COVID-19 pathogenesis and act as antifibrosis drugs. From the screened-out chemical candidates, we used the PADEL descriptors to extract both the PubChem and substructure fingerprints from the compounds, which are the representations for certain morphological properties like functional groups, molecular bonds, and other physiological properties that influence the overall activity of the chemicals in the biological environment. From the mapped database of chemicals and their respective phenotypes, we hypothesized that connecting the mentioned chemicals against the phenotypes might provide us with an idea regarding the mechanistic action of these compounds to treat the phenotypes, which were also the symptoms produced during COVID-19 and related to CRS activity. Hence, we decided to cluster these chemicals, and more specifically, the molecular fingerprints, against the phenotypes to determine which specific fingerprints were mostly correlating with the phenotypes. This would provide us with a conclusion regarding the morphological properties of the chemical, which are actively involved in the biological activity against the phenotype and much more mechanistically involved. Therefore, this approach could lead to the determination of various scaffolds, which are only fingerprints and could be used as an integral factor in the design of bioactive compounds during drug discovery methodologies. The representation of the hierarchical cluster with a dendrogram to correlate the fingerprints against the phenotypes is given below in the following figure (Fig. 8) (Supplementary S9). Additionally, we also provided a tabular representation (Table 2) of the molecular fingerprints that were used in the dendrogram against the expressed phenotypes and their respective annotations regarding the SMILES presentation and its morphological meaning, which included the chemical bonds and structural information like scaffolds and functional groups that constitute the formation of the respective compound. Moreover, information regarding their intensity of correlation is also provided as per the color code obtained from the heatmap to determine its presence along the phenotypes. Moreover, from the CTD server, we performed another set of analyses, which gave us the results regarding the upregulation and downregulation levels of the chemicals against their respective genes. The main idea of this process was to deduce the transcriptomic-level information of the genes and their interactions with the screened chemicals to influence their biological expression within the localized tissues and organs. Hence, from the results we obtained, more specific methodologies can be carried out to screen the specific chemicals that would contribute to downregulating certain genes that may be involved in overexpression due to the pathogenesis of COVID-19 and PF, and vice versa. Therefore, by studying the expression analysis across the chemicals and their respective genes, we can put more focus into developing chemical candidates, which might be an important factor in controlling gene expression within the organs or tissues where CRS activity would be the highest, hence enabling us to treat the biological conditions that involve abnormal gene

expression levels leading to the pathogenesis of PF. The data containing the gene expression information for chemicals against their respective genes is attached to the supplementary material and consists of information regarding the chemicals and the respective genes they interacted with, along with information revolving around upregulation or downregulation of the gene expression within its biological pathways.

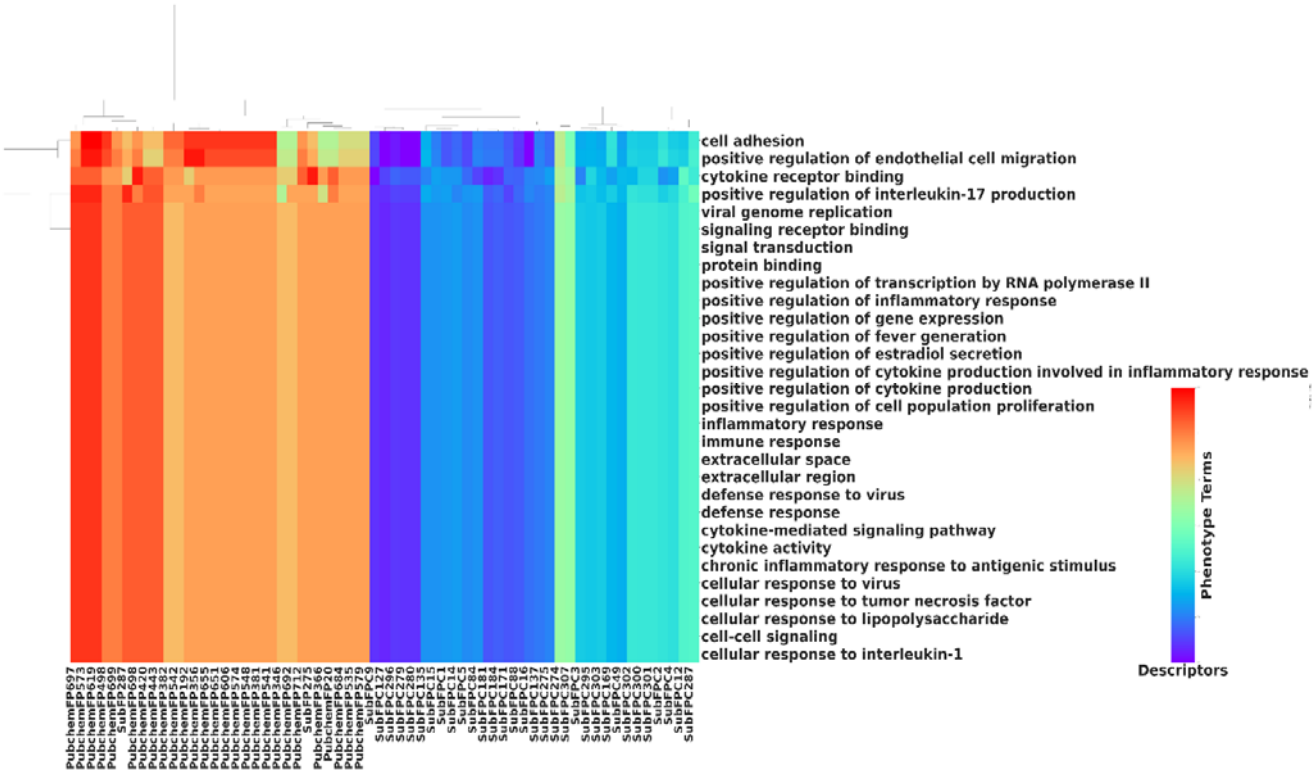


Fig.8: Diagrammatic representation of Hierarchical Clustering map of chemicals' fingerprints against the mapped phenotypes of COVID, CRS, and PF.

Table 2: Tabular representation of the molecular fingerprints and their annotated terminologies with the level of correlation as per the heatmap.

PADEL Fingerprint Names	Fingerprint Annotation (SMILES / Description)	Correlation Intensity
PUBCHEMFP697	C-C-C-C-C-C©-C	High
PUBCHEMFP619	O-C-C=C-C	High
PUBCHEMFP698	O-C-C-C-C-C-C-C	High
PUBCHEMFP443	C(-C)(=O)	High
PUBCHEMFP542	O-C:C-[#1]	High
PUBCHEMFP356	C(~C)(:C)(:C)	High
PUBCHEMFP651	O-C:C:C-C	High
PUBCHEMFP574	O-C-C:C-C	High
PUBCHEMFP381	C(~O)(:C)	High
PUBCHEMFP346	C(~C)(~H)(~O)	High

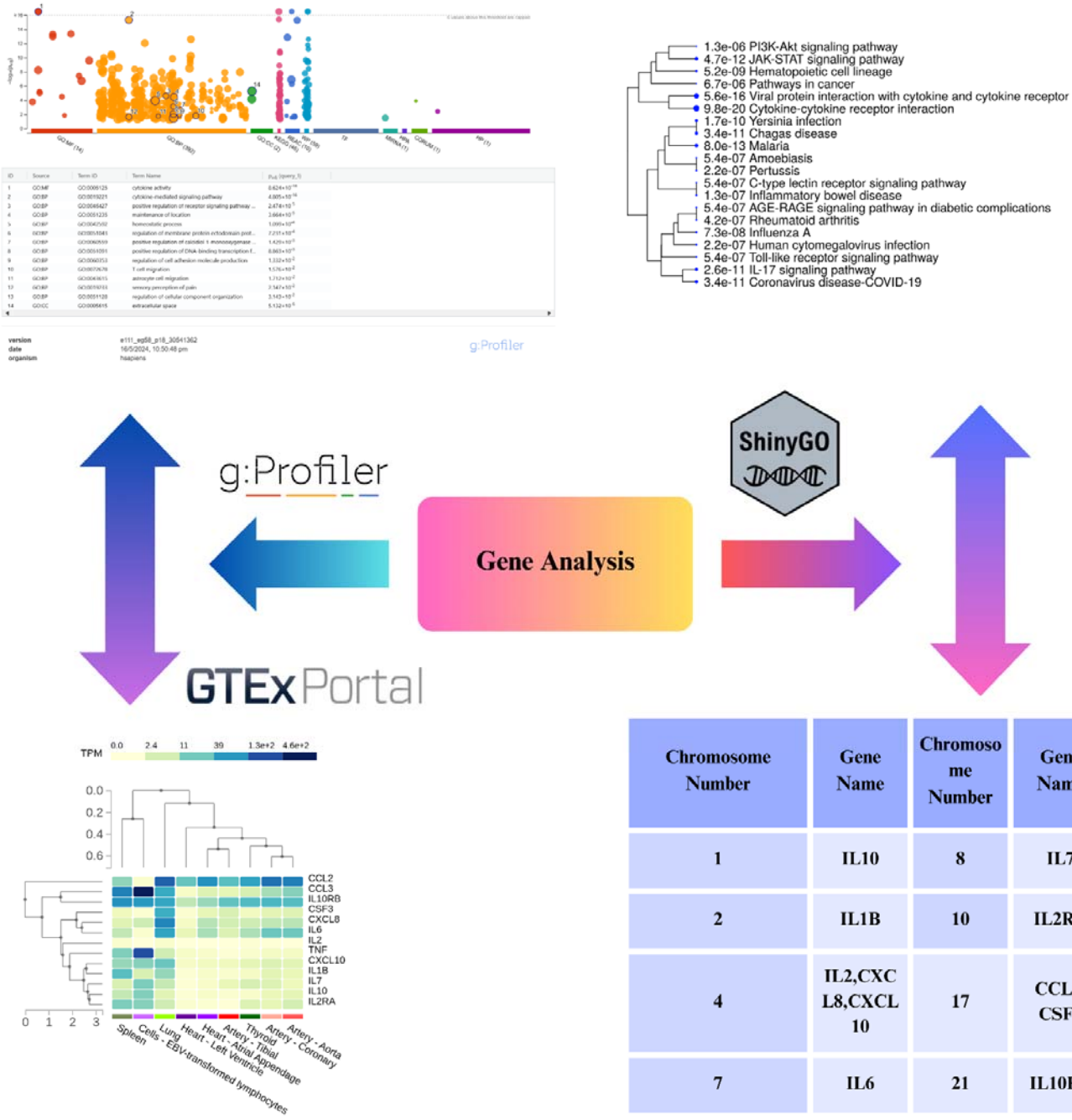
PUBCHEMFP712	C-C©-C©-C	High
PUBCHEMFP366	C(~H)(~O)	High
PUBCHEMFP684	O=C-C-C-C-C	High
PUBCHEMFP579	O=C-C-C-C	High
SUBPFC127	Peptide middle	Low
SUBPFC279	Annelated rings	Low
SUBPFC135	Vinylogous carbonyl or carboxyl derivative	Low
SUBPFC1	Primary carbon	Medium
SUBPFC5	Alkene	Medium
SUBPFC181	Hetero N nonbasic	Medium
SUBPFC171	Arylchloride	Medium
SUBPFC16	Dialkylether	Low
SUBPFC275	Heterocyclic	High
SUBPFC307	Chiral center specified	High
SUBPFC295	C ONS bond	Medium
SUBPFC169	Phenol	Medium
SUBPFC302	Rotatable bond	Medium
SUBPFC301	1,5-Tautomerizable	Medium
SUBPFC4	Quaternary carbon	Medium
SUBPFC287	Conjugated double bond	Medium

From the above figure, we observed the formation of clusters along the PubChem fingerprints and the Substructure fingerprints against the cell adhesion and cytokine-related signaling phenotypes respectively. This could provide a conclusion that the specific group of molecular fingerprints for each chemical candidate is significantly correlated with the expression of the mentioned phenotypes. The obtained results and the tabular representation of the descriptor information also allow us to infer the relationships and biological connections among these specific descriptors and would contribute towards performing further studies on these specific morphological factors of the chemicals to analyze the mechanistic action and pathways they are involved into to provide responses to control or inhibit the immunological reactions expressed by the mentioned phenotypes acting as potential drug candidates which could be improvised to inhibit the negative feedback produced by abnormal cytokine release which relates to the pathogenesis of PF within COVID-19 patients.

3.5 Analysis of genes obtained from the JI cluster map against the phenotypes to specify their biological action and involvement in the overall pathogenesis of COVID-19.

From the list of genes (Supplementary S10) obtained from the results of the JI correlation map across gene pairs against the various pathways in Section 3.4, we decided to perform a series of analyses using the methods described in Section 2.6 to calculate the functional enrichment parameters along with the gene ontology (GO) terms to determine the general molecular functions, biological properties, and cellular

localisations that are shared by the mentioned list of genes. These properties would help us determine the biological significance of these genes within the overall functioning of the immune system and regarding their involvement in the progression of COVID-19 and how they are related to the development of PF symptoms through CRS actions. Additionally, we also obtained the COVID-19 pathway along with the genes that are mentioned and marked with red to indicate their presence. Further data was collected by us, which included the chromosomal locations of the genes within the human genome along with the organs and their tissues where the gene expression activity is significantly localized and involved in their biological pathways. Hence, the results obtained from this analysis further verified the discussions regarding the connection between CRS and PF pathogenesis within COVID-19 patients and enabled us to connect the genes and the biological pathways that are shared by the above three and perform a detailed study about the respective phenotypes expressed along their progression. The following diagrams (Fig. 9) represent the gene analysis results obtained from our analysis. Additionally, we also obtained the single cell expression analysis for the selected genes to understand their significant sites of differential expression (Supplementary S11 Fig 3s). We also present the biological pathway of COVID involving the mentioned top analyzed genes along with the statistical inference of the respective GO terminologies like Molecular Functions, Biological Processes, and Cellular Components with their respective adjusted P-values to understand how significant they are within the pathogenesis of COVID-19 (Supplementary S11 Fig 4s-8s).



Chromosome Number	Gene Name	Chromosome Number	Gene Name
1	IL10	8	IL7
2	IL1B	10	IL2RA
4	IL2,CXCL8,CXCL10	17	CCL2, CSF3
7	IL6	21	IL10RB

Fig. 9: Analysis of obtained genes to gather their biological data consisting of GO terms, chromosomal locations, organ bases expression, and phylogenetic properties.

The graphical representation (Supplementary S11 Fig 3s) included the results obtained from single-cell expression analysis for the top-performing list of genes within a wide spectrum of tissues and cells of various organs they are differentially expressed within. The differential expression facilitated by the gene activity could be due to cytokine release progression, which is the main aim of our manuscript, or to the presence of an infection or immune response through antigen activity against the receptor binding processes. As we are more focused on studying the effects of cytokines and their responsible gene constituents, which influence the CRS within COVID-19 patients and contribute to the overall progression of PF phenotypes, we mainly observe the expression levels within the tissues and cells that belong to the morphology of the lung and respiratory system. From the above results, we observed that, besides the significant expression of all genes, the maximum expressions within the lung tissues are mainly characterized by the interleukin proteins, more specifically IL10RB, IL2RA, and IL7, respectively, which gave us a clear indication that the interleukin expression, which is a major factor responsible for cytokine storms and inflammatory responses induced by CRS, was being actively expressed within the lung tissues, hence influencing the initiation of the immune response pathways along with the respective occurrence of the phenotypes within COVID-19 patients.

Therefore, from the provided results, we could understand how the list of obtained genes, which contain the most biological pathways associated with COVID-19, CRS, and PF, are statistically significant in terms of GO terminologies along with their phylogenetic relationships across related biological pathways where they are significantly expressed. Moreover, we also observed the chromosomal locations where the genes are most likely to be found, along with the organs and tissues where they are significantly expressed. The presence of genes along the lung and its tissues denotes the involvement of a series of immune reactions and inflammatory responses, most likely due to the action of cytokines through CRS, which, on abnormal levels, cause excessive infections leading to fibrosis along the tissues, hence the progression of PF.

3.6 Development of QSAR Models to create predictive models of potent chemical candidates to target the abnormal expression of TNF and IL6 genes due to their major involvement in initiating CRS symptoms within COVID patients.

As discussed previously in Section 2.7, we obtained the bioactivity dataset represented by the IC50 (inhibition concentration at 50%) values of various chemicals and drug candidates, which correspond to the target proteins TNF and IL6, respectively. These chemical candidates and samples are obtained separately from the CHEMBL database and filtered to only contain the IC50 values, marking compounds as active if their potency value lies within the range of less than 1000 nM and labeling compounds inactive if their value exceeds 10000 nM. From the data collection step, we observed the dataset to be much bigger in the case of TNF, containing a wide range of bioactivity values, whereas a smaller and limited dataset with only lower IC50 values in the case of IL6, which were entirely categorized into the active bioactivity state. Hence, to

process the larger dataset of TNF inhibitors, we deployed both a classification ML model using a Random Forest Classifier (RF) and a regression model using a genetic algorithm coupled with multilinear regression (GA-MLR). RF utilizes the concept of a decision tree to calculate every molecular feature for their biological activity and clusters the molecules based on their bioactivity, classifying them into active and inactive states. Moreover, RF also provides a feature importance plot, which provides us with information about the set of selected molecular fingerprints, which significantly played an integral part in the overall model prediction process. GA-MLR is a robust regression model that is extremely useful in handling complex datasets with a high number of features and uses the concept of natural selection through generational data processing to filter out the best features contributing to the maximum positive performance of the training process of the model. However, while obtaining and processing the chemical dataset for IL6, we observed that the number of samples was too low to be used in the classification modeling and that the bioactivity values only belonged to the active class. Henceforth, we deployed the GA-MLR model to create a regression algorithm to predict the bioactivity values of IL6 inhibitors from the trained data provided. GA-MLR, as mentioned previously, utilizes a similar concept of natural selection and generational feature processing like genetic processes, where from a set of given data points, the algorithm creates all possible combinations of the features against their respective biological activity as targets to obtain the best features that overall contribute to the net performance of the prediction model. Hence, observing the regression metrics like mean absolute error (MAE) and coefficient of determination (R2) scores, we select the best models to further analyze the graphical validations, specifically the regression plot and the applicability domain (AD), to filter out the best chemicals that were predicted from both training and testing partitions. Additionally, we also concluded that the presence of the selected molecular descriptors, which were denoted as most significant while training the model, was the most important morphological characteristic, whose alterations and presence were directly proportional to their entire mechanism of action against the selected biological pathways and gene expression responsible for COVID-19 pathogenesis, leading to the progression of PF through CRS development. The results corresponding to the model performance and analysis of chemicals from the model to verify its involvement within the biological pathways and phenotypes corresponding to the pathogenesis of COVID-19 and PF progression are provided in the following diagrams (Figure 10-11).

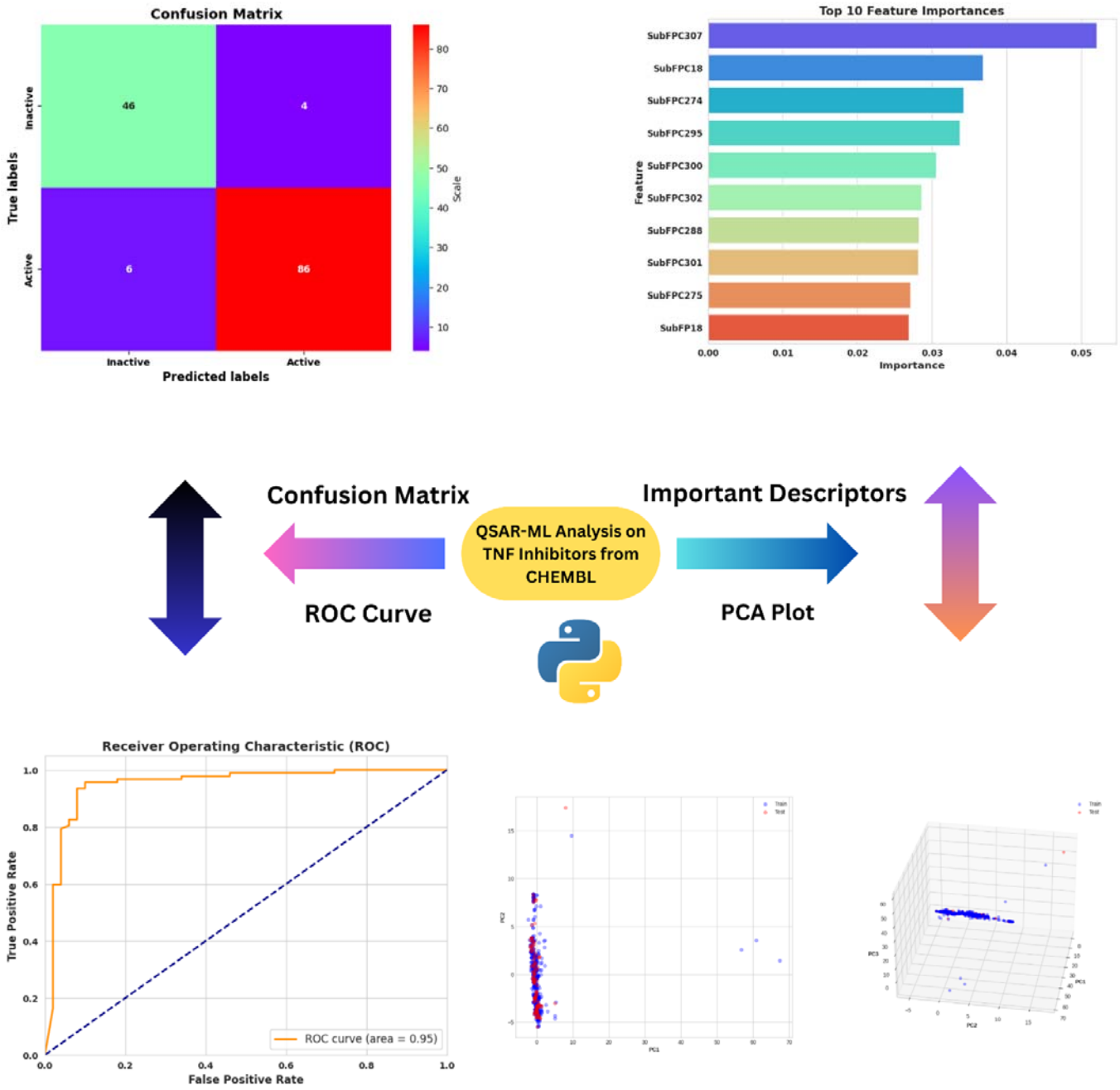


Fig. 10: QSAR ML modeling against TNF inhibitors using Random Forest Classifier

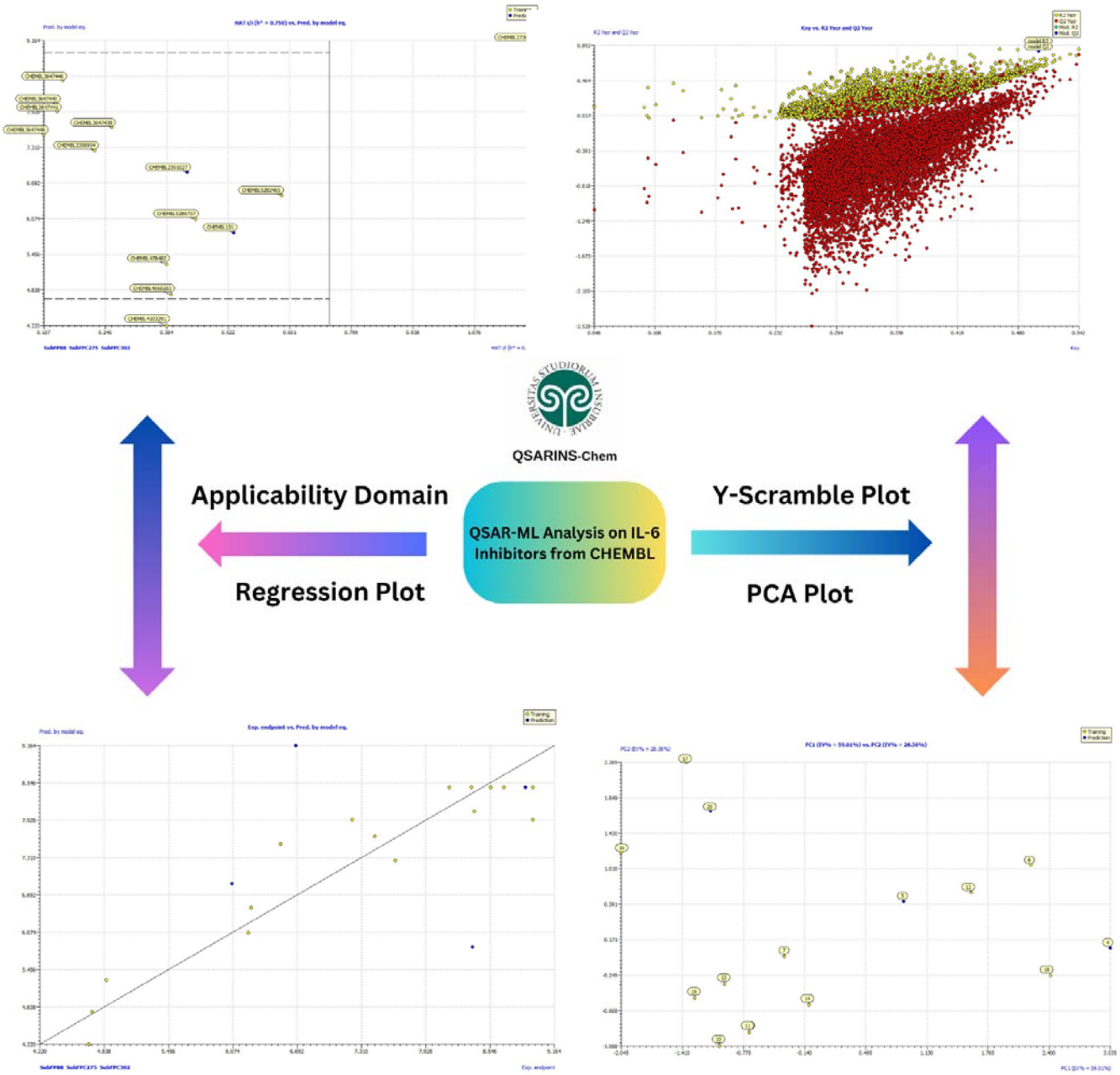


Fig. 11: QSAR ML modeling against IL-6 inhibitors using GA-MLR

From the following representation, we can understand the overall dynamics associated with the mechanistic mode of action across the biological pathways and genes the chemicals are mainly associated with. From the feature importance plot generated from the RF algorithm and GA-MLR, we obtained a set of specific molecular fingerprints that influence the overall prediction model more significantly than the others. Hence, from this, we could deduce that these specific descriptors represent certain morphological and physiological parameters among the chemicals whose presence is highly influential within the biological pathways to exhibit a specific expression activity of target genes, therefore leading to a mechanistic mode of action to either promote the development of CRS into PF within COVID patients or provide therapeutic effects against the detrimental effects caused by the viral progression by the Coronavirus, which leads to the

progression of CRS leading to PF. Additionally, we obtain the top target genes that are associated with these chemical candidates, which in this case mainly correspond to TNF or IL-6, respectively, as mentioned above. Obtaining the top predicted chemicals from the deployed algorithms, we analyze their target genes to ensure their biological activity is significantly concentrated around TNF or IL-6 as per the model, respectively. Once the chemical candidate is properly analyzed with its chemical-target interaction, we map them with our dataset containing the biological pathways and expressed phenotypes associated with the pathogenesis of COVID-19 and its progression into PF through the development of CRS and obtain a graphical network that is visualized by the respective target gene at the center and their respective biological pathways and phenotypes they would possibly interact with through their mechanistic methodology. By utilizing this procedure, we could connect the QSAR result with the network biology part to verify the mode of action of the predicted chemicals from our ML models and analyze their interactions and biological involvement with the pathways and phenotypes associated with our study. Additionally, we analyze the model performance metrics to provide concrete evidence regarding the overall performance of the deployed ML models. The main metrics we targeted for RF and GA-MLR included the confusion matrix, important feature analysis, ROC plot, applicability domain, and regression plot, respectively. The confusion matrix derived from the RF classifier applied to the TNF inhibitor dataset provided us with the accuracy of classification between the active and inactive chemicals based on their pIC50 values. Additionally, the Important Features graph provided us with the most significant molecular fingerprint descriptors for each chemical within the dataset that overall influence the model performance and whose presence is directly connected to the biological activity exerted by the chemicals along the biological pathways associated with COVID-19 and PF progression. The ROC plot enables us to verify the classification accuracy across the true positive prediction of samples, which denotes a perfect model if the area covered by the plot is higher, corresponding with the R2 score. The Applicability Domain and Regression Plot were specifically obtained from the GA-MLR algorithm applied to the IL6 inhibitor dataset, which only had chemicals with potent inhibitory values and enabled us to determine the chemical candidates from the dataset that were properly predicted and responded to a particular statistical threshold that determined its performance levels within the prediction system as well as establish a compact linear relationship between the significant molecular fingerprints and their respective inhibitory concentration values (pIC50). Hence, by observing the following representations, we determined the accuracy and performance levels of the models we developed, which is necessary to analyze the mechanism of action of the respective chemicals within the biological pathways associated with the pathogenesis of COVID-19 and the phenotypes that were expressed along the process that corresponded to the development of PF through CRS progression. We also included the model performance metrics with the above results. Additionally, we also discussed in the previous section regarding analyzing the target genes from the selected chemicals to develop a complex network to visualize their connection against the phenotypes and pathways associated with the pathogenesis of COVID-19 leading to the development of CRS leading to PF which we have included within the supplementary file (Supplementary S11 Fig 9s-10s)

4. Conclusion

Therefore, concluding the manuscript, we discussed the work related to determining the biological connections shared between the pathogenesis of COVID-19 and the mechanism of cytokine proteins that respond to providing inflammatory responses to the viral pathways. Moreover, we also established the connection between PF progression and CRS in COVID patients and visualized the respective phenotypes and biological pathways that are responsible for the development of the severity of the symptoms expressed by them by implementing knowledge graph-based hierarchical clustering matrix representations. This overall study enabled us to understand the mechanistic action of the genes and their significance in the initiation of complex immune responses and inflammatory actions across sites of infection, which are the backbone of understanding the complexity of fibrosis mechanisms. On the other hand, we determined the actions of various bioactive compounds and substances that overall contribute to or inhibit the immunological processes that were expressed in the form of certain phenotypes like protein binding and inflammation, which are mainly observed in the biological conditions associated with abnormal cytokine release connected to PF. Additionally, we performed a well-detailed analysis within the chemical morphology to study the specific molecular descriptors associated with the structural map of the chemical and their biological involvement across their respective biological pathways and phenotypes which provided us with positive feedback and idea into the discovery of drugs and potent bioactive compounds which would have the potential to combat the severe infections caused by the irregularities in cytokine release within COVID patients. Ultimately, from the list of genes that were commonly participating across COVID, CRS, and PF throughout their pathogenesis, we were also able to extract the much more specific list of genes among them that were more significantly expressed with the biological progression of COVID and actively participating along the pathways related to infections and fibrosis within lung tissues due to the abnormal release of cytokines from CRS mechanisms.

We also elucidated the mechanism of action of toxicological and therapeutic chemicals using multi-dimensional knowledge graph-based hierarchical clustering matrix representations by understanding their modulation at the genetic level through pathways and phenotype alterations. Additionally, we also developed robust quantitative structural activity relation (QSAR) models in combination with system biology approaches against the top expressed hub genes (TNF and IL-6) of our knowledge graph to understand the key pharmacophores of biologically active small molecules that can be further implemented to develop drugs against COVID-19. Our established multi-dimensional knowledge graph-based hierarchical clustering matrix thereby demonstrated a novel overview of the significant insights into the molecular mechanism elucidation of COVID-PF-CRS tri-sectional diseased condition by highlighting the clusters of deregulated genes based on commonalities in the pathways modulated and phenotypes altered; thereby identifying the best biomarkers for designing new drug molecules. Additionally, our established clustering algorithms along with the QSAR models also highlight the mechanism of action and physicochemical

attributes required for designing new drug candidates to treat COVID-PF-CRS tri-sectional disease condition. Our findings offer promising insights for the development of the next generation of anti-COVID therapies, as paving the way for the design of molecules that target not only the viral proteome but also enhance the host immune response, thus providing a comprehensive and effective approach to combatting COVID-19.

Acknowledgments: The authors extend their sincere appreciation to the following organizations for their invaluable support and funding: 1) the Finnish Cultural Foundation; 2) the Tampere Tuberculosis Foundation (Ashok Aspatwar); 3) the Academy of Finland; and 4) the Jane and Aatos Erkko Foundation (Seppo Parkkila).

Conflicts of interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

5. References

[1] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 2020;382:727–33. https://doi.org/10.1056/NEJMoa2001017.

[2] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet (London, England) 2020;395:565–74. https://doi.org/10.1016/S0140-6736(20)30251-8.

[3] Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. Eur J Epidemiol 2020;35:1123–38. https://doi.org/10.1007/s10654-020-00698-1.

[4] Ioannidis JPA. Infection fatality rate of COVID-19 inferred from seroprevalence data. Bull World Health Organ 2021;99:19-33F. https://doi.org/10.2471/BLT.20.265892.

[5] Welch C. Age and frailty are independently associated with increased COVID-19 mortality and increased care needs in survivors: results of an international multi-centre study. Age Ageing

2021;50:617–30. https://doi.org/10.1093/ageing/afab026.

[6] Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis 2020;20:669–77. https://doi.org/10.1016/S1473-3099(20)30243-7.

[7] Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020;584:430–6. https://doi.org/10.1038/s41586-020-2521-4.

[8] Zhang Y, Geng X, Tan Y, Li Q, Xu C, Xu J, et al. New understanding of the damage of SARS-CoV-2 infection outside the respiratory system. Biomed Pharmacother 2020;127:110195. https://doi.org/https://doi.org/10.1016/j.biopha.2020.110195.

[9] Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature 2020;581:465–9. https://doi.org/10.1038/s41586-020-2196-x.

[10] Yu X, Sun S, Shi Y, Wang H, Zhao R, Sheng J. SARS-CoV-2 viral load in sputum correlates with risk of COVID-19 progression. Crit Care 2020;24:170. https://doi.org/10.1186/s13054-020-02893-8.

[11] Wang Y, Zhang L, Sang L, Ye F, Ruan S, Zhong B, et al. Kinetics of viral load and antibody response in relation to COVID-19 severity. J Clin Invest 2020;130:5235–44. https://doi.org/10.1172/JCI138759.

[12] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet (London, England) 2020;395:507–13. https://doi.org/10.1016/S0140-6736(20)30211-7.

[13] Guan W-J, Ni Z-Y, Hu Y, Liang W-H, Ou C-Q, He J-X, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020;382:1708–20. https://doi.org/10.1056/NEJMoa2002032.

[14] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England) 2020;395:497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

[15] Polak SB, Van Gool IC, Cohen D, von der Thüsen JH, van Paassen J. A systematic review of pathological findings in COVID-19: a pathophysiological timeline and possible mechanisms of disease progression. Mod Pathol an Off J United States Can Acad Pathol Inc 2020;33:2128–38. https://doi.org/10.1038/s41379-020-0603-3.

[16] Ferrara JL, Abhyankar S, Gilliland DG. Cytokine storm of graft-versus-host disease: a critical effector role for interleukin-1. Transplant Proc 1993;25:1216–7.

[17] Bassetti M, Vena A, Giacobbe DR. The novel Chinese coronavirus (2019-nCoV) infections: Challenges for fighting the storm. Eur J Clin Invest 2020;50:e13209. https://doi.org/10.1111/eci.13209.

[18] Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir Med 2020;8:e21. https://doi.org/10.1016/S2213-2600(20)30116-8.

[19] Mahmud-Al-Rafat A, Majumder A, Taufiqur Rahman KM, Mahedi Hasan AM, Didarul Islam KM, Taylor-Robinson AW, et al. Decoding the enigma of antiviral crisis: Does one target molecule regulate all? Cytokine 2019;115:13–23. https://doi.org/10.1016/j.cyto.2018.12.008.

[20] Vancheri C, Failla M, Crimi N, Raghu G. Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology. Eur Respir J 2010;35:496–504. https://doi.org/10.1183/09031936.00077309.

[21] Richeldi L, du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. N Engl J Med 2014;370:2071–82. https://doi.org/10.1056/NEJMoa1402584.

[22] Lancaster L, Crestani B, Hernandez P, Inoue Y, Wachtlin D, Loaiza L, et al. Safety and survival data in patients with idiopathic pulmonary fibrosis treated with nintedanib: pooled data from six clinical trials. BMJ Open Respir Res 2019;6:e000397. https://doi.org/10.1136/bmjresp-2018-000397.

[23] King TEJ, Bradford WZ, Castro-Bernardini S, Fagan EA, Glaspole I, Glassberg MK, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. N Engl J Med 2014;370:2083–92. https://doi.org/10.1056/NEJMoa1402582.

[24] Liang J, Zhang Y, Xie T, Liu N, Chen H, Geng Y, et al. Hyaluronan and TLR4 promote surfactant-protein-C-positive alveolar progenitor cell renewal and prevent severe pulmonary fibrosis in mice. Nat Med 2016;22:1285–93. https://doi.org/10.1038/nm.4192.

[25] Schruf E, Schroeder V, Le HQ, Schönberger T, Raedel D, Stewart EL, et al. Recapitulating idiopathic pulmonary fibrosis related alveolar epithelial dysfunction in a human iPSC-derived air-liquid interface model. FASEB J Off Publ Fed Am Soc Exp Biol 2020;34:7825–46. https://doi.org/10.1096/fj.201902926R.

[26] Barbas-Filho J V, Ferreira MA, Sesso A, Kairalla RA, Carvalho CR, Capelozzi VL. Evidence of type II pneumocyte apoptosis in the pathogenesis of idiopathic pulmonary fibrosis (IFP)/usual interstitial pneumonia (UIP). J Clin Pathol 2001;54:132–8. https://doi.org/10.1136/jcp.54.2.132.

[27] Tatler AL, Goodwin AT, Gbolahan O, Saini G, Porte J, John AE, et al. Amplification of TGFβ Induced ITGB6 Gene Transcription May Promote Pulmonary Fibrosis. PLoS One

2016;11:e0158047. https://doi.org/10.1371/journal.pone.0158047.

[28] Tatler AL, Habgood A, Porte J, John AE, Stavrou A, Hodge EK, et al. Reduced Ets Domain-containing Protein Elk1 Promotes Pulmonary Fibrosis via Increased Integrin $α$v$β$6 Expression*. J Biol Chem 2016;291:9540–53.

[29] Antoniades HN, Bravo MA, Avila RE, Galanopoulos T, Neville-Golden J, Maxwell M, et al. Platelet-derived growth factor in idiopathic pulmonary fibrosis. J Clin Invest 1990;86:1055–64. https://doi.org/10.1172/JCI114808.

[30] Abdollahi A, Li M, Ping G, Plathow C, Domhan S, Kiessling F, et al. Inhibition of platelet-derived growth factor signaling attenuates pulmonary fibrosis. J Exp Med 2005;201:925–35. https://doi.org/10.1084/jem.20041393.

[31] Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res 2023;51:D1257–62. https://doi.org/10.1093/nar/gkac833.

[32] OLIVEROS J. C. VENNY. An interactive tool for comparing lists with Venn diagrams. Http://BioinfogpCnbCsicEs/Tools/Venny/IndexHtml 2007.

[33] G. van Rossum. Python tutorial, Technical Report CS-R9526. Cent Voor Wiskd En Inform (CWI), 1995.

[34] McKinney W. Data Structures for Statistical Computing in Python. SciPy, 2010.

[35] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504. https://doi.org/10.1101/gr.1239303.

[36] Hagberg A, Swart P, Chult D. Exploring Network Structure, Dynamics, and Function Using NetworkX. 2008.

[37] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32:1466–74. https://doi.org/10.1002/jcc.21707.

[38] Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res 2023;51:W207–12. https://doi.org/10.1093/nar/gkad347.

[39] Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 2020;36:2628–9. https://doi.org/10.1093/bioinformatics/btz931.

[40] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653.

[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. ArXiv 2011;abs/1201.0.

[42] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 2013;34:2121–32.