國立臺灣大學電機資訊學院生醫電子與資訊學研究所

博士論文

Graduate Institute of Biomedical Electronics and Bioinformatics
College of Electrical Engineering and Computer Science
National Taiwan University
Doctoral Dissertation

粒線體與多重器官損傷在COVID-19的相關性之轉錄組和機器學習分析

Transcriptome and machine learning analyses of the correlation between mitochondria and multiorgan damage in COVID-19

張裕宇
Yu-Yu Chang

指導教授：魏安祺 博士
Advisors: An-Chi Wei, Ph.D.

中華民國113年8月
August 2024

doi:10.6342/NTU202402334

誌謝

非常感謝我的指導教授魏安祺博士的鼓勵、支持和指導，才讓我得以順利完成這篇博士論文。在這段艱辛而充滿挑戰的旅程中，老師給予了我許多的啟發和幫助，除了在學術研究上提供了寶貴的建議和指導，還在我面臨困難和瓶頸時給予了我信心和支持。老師的寬容、耐心和知識讓我在學術道路上受益良多，這些經驗和教誨將成為我未來研究中不可或缺的一部分。

另外也感謝我的口試委員林致廷教授、阮雪芬教授、陳沛隆教授以及陳倩瑜教授在博士論文計畫審查及口試的期間給予的寶貴建議和指導。老師們的專業見解和細心指導，幫助我改進與提升了研究方向與內容，為此我深表感謝。

最後，我要感謝在博士就學期間指導過我的每位老師和幫助過我的每一個人。你們的教導、支持和楷模使我受益匪淺，這些經驗和教誨將伴隨我終生。我會永遠銘記在心，感謝你們的幫助和悉心指導。

doi:10.6342/NTU202402334

中文摘要

許多研究表明，嚴重急性呼吸道症候群冠狀病毒 2 型(SARS-CoV-2)可以透過多種方式損害多個器官，包括透過血管收縮素轉化酶 2 (ACE2)、促炎性細胞因子風暴或其他次級途徑促進的直接病毒入侵。而 long COVID 是指在初次感染 COVID-19 後，個體經歷持續的器官損傷或出現新的症狀。

敗血症引起多重器官損傷的機制包括引發全身性過度發炎反應、免疫系統過度活化、影響細胞能量代謝等幾個重要面向。這些機制相互作用，共同導致多重器官損傷。

本研究利用 Gene Expression Omnibus (GEO) 資料庫的公開轉錄組資料來識別在 COVID-19、敗血症和其他急性呼吸道感染疾病中表達顯著差異的基因。進一步的研究檢視與敗血症、其他急性呼吸道感染疾病以及 SARS-CoV-2 誘導的粒線體、心臟、肝臟和腎臟損傷相關的途徑。接著對顯著差異表達的基因進行挑選和排序，並使用特徵重要性對具有生物途徑意義的基因進行排序，作為機器學習驗證的特徵。

透過效能、樣本大小、不平衡資料狀態和過度擬合來評估機器學習的樣本集選擇。機器學習也透過調整基因清單來幫助評估生物途徑的假設。隨後進行的深入研究檢視了基因和相關途徑，以了解粒線體與多重器官損傷在 COVID-19 的關聯。

研究結果表明，ACE2、促炎性細胞因子風暴和線粒體損傷對 COVID-19 導致的多器官損傷存在著關聯。敗血症和 COVID-19 引起的多重器官損傷的機制不同。

doi:10.6342/NTU202402334

而且，粒線體損傷是導致 long COVID 的關鍵因素之一。這些發現可以做為研究潛在的醫療治療以應對 SARS-CoV-2 引起的多器官損傷。

關鍵字：SARS-CoV-2、COVID-19、RNA-Seq、急性呼吸道感染疾病、敗血症、轉錄組分析、機器學習、粒線體、基因表現、路徑分析、交叉驗證、過度擬合

doi:10.6342/NTU202402334

Abstract

The extensive research has shown that Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) can harm several organs through various means, including direct viral invasion facilitated by angiotensin-converting enzyme 2 (ACE2), inflammatory cytokine storms, or other secondary pathways. Long COVID is when individuals experience persistent, lasting damage to organs or new symptoms for several weeks or months after recovering from an initial COVID-19 infection.

The mechanisms by which sepsis causes multiple organ damage include several vital aspects, for example, triggering a systemic excessive inflammatory response, the excessive activation of the immune system, and affecting cellular energy metabolism. These mechanisms interact with each other, collectively leading to multiorgan damage.

Publicly available transcriptome data from the Gene Expression Omnibus (GEO) database was utilized to identify genes significantly differently expressed in COVID-19, sepsis, and other non-COVID-19 acute respiratory infections. A further investigation examined pathways connected to sepsis, other non-COVID-19 acute respiratory infections, and SARS-CoV-2-induced mitochondrial, cardiac, hepatic, and renal damage. Statistical methods were used to identify and rank significantly differentially expressed genes, and feature importance was used for rating biologically significant genes as machine learning verification features.

The sample set selection for machine learning was evaluated through performance, sample size, imbalanced data state, and overfitting. Machine learning also assisted in evaluating biological hypotheses by adjusting gene lists. A subsequently thorough study examined genes and pathways to figure out the correlation between mitochondria and multi-organ damage in COVID-19.

doi:10.6342/NTU202402334

The research findings suggest a link between ACE2, inflammatory cytokine storms, and mitochondrial damage in COVID-19, potentially contributing to multiorgan damage. The mechanism of multiorgan damage caused by sepsis and COVID-19 is different. Moreover, mitochondrial damage is one of the critical factors leading to long COVID. These findings indicate that potential medical treatments could be studied to address the damage to multiple organs caused by SARS-CoV-2.

Keywords: SARS-CoV-2, COVID-19, RNA-Seq, acute respiratory infections, sepsis, transcriptome analysis, machine learning, mitochondria, gene expression, pathway analysis, cross validation, overfitting

doi:10.6342/NTU202402334

Contents

doi:10.6342/NTU202402334

doi:10.6342/NTU202402334

doi:10.6342/NTU202402334

List of Figures

doi:10.6342/NTU202402334

List of Tables

doi:10.6342/NTU202402334

Chapter 1 Introduction

Part of the thesis is from Chang YY et al., "Transcriptome and machine learning analysis of the impact of COVID-19 on mitochondria and multiorgan damage" PLoS One (2024). Multiple organ damage refers to the simultaneous dysfunction or failure of two or more vital organs under severe clinical conditions, such as severe infections, trauma, or post-surgical complications. This condition can involve various organ systems, including the respiratory, cardiovascular, kidneys, liver, nervous, and digestive systems, among others. When these organs fail to function properly, the patient's overall condition can rapidly deteriorate, necessitating swift and comprehensive medical intervention. Multiple organ damage is common in COVID-19, sepsis, and other related diseases caused by viral or bacterial infections (1).

The causes of multiple organ damage are diverse, including systemic inflammatory responses triggered by severe infections, tissue damage during ischemia and reperfusion, an overactive or dysregulated immune system, and direct impacts from certain drugs or toxins (2).

Patients afflicted with the coronavirus disease 2019 (COVID-19) frequently experience respiratory complications, which can subsequently lead to the development of acute respiratory distress syndrome (ARDS). This syndrome, brought on by COVID-19 pneumonia, is the primary cause of mortality and prolonged lung damage in patients. While the respiratory system is the most commonly impacted in clinical cases triggered by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), it is essential to note that the virus can potentially affect any organ in the body. A significant proportion of individuals who succumbed to COVID-19 exhibited respiratory system impairment, and approximately half experienced multi-organ damage(3). It is common for critically

doi:10.6342/NTU202402334

ill patients to experience the involvement of multiple organs (4). It has been observed that many patients with COVID-19 exhibit symptoms beyond the classical respiratory distress. These patients may also experience systemic symptoms, such as cardiovascular, hepatic, or renal failure, and coagulation disorders. Various studies have indicated that organ damage can occur in the lungs (33% of patients), heart (32%), kidneys (12%), liver (10%), pancreas (17%), and spleen (6%). Furthermore, 66% of study participants displayed damage to one or more organ systems, while 25% of patients exhibited multiorgan damage with varying degrees of overlap between different organs (5). The presence of such complications is indicative of the multi-organ involvement of the disease. Therefore, a comprehensive assessment of these symptoms and the prompt implementation of targeted therapeutic interventions may prove crucial in mitigating the adverse outcomes associated with COVID-19.

1.1 Multiorgan damage by SARS-CoV-2

The study on cardiac SARS-CoV-2 infection has identified specific pro-inflammatory transcriptomic changes in the hearts of patients infected with the virus. The study found that the upregulated pro-inflammatory genes mostly originated from endothelial cells present in the cardiac tissue. These findings suggest that SARS-CoV-2 infection in the heart triggers a pro-inflammatory response at the transcriptomic level, with endothelial cells being the primary target for viral infection within the cardiac tissue. The pro-inflammatory response in the heart may potentially contribute to developing or worsening heart failure (6). Numerous mechanisms have been proposed to explain the potential for kidney injury during SARS-CoV-2 infection, including direct cytopathic effects on renal tissue, cytokine-mediated damage, and the impact of

doi:10.6342/NTU202402334

the inflammatory phase on renal function. Research has shown that certain COVID-19 patients, including those without pre-existing renal pathologies, may exhibit indications of kidney injury, with over 30% of inpatients developing renal dysfunction. The interaction between SARS-CoV-2 and the renin-angiotensin-aldosterone system (RAAS) through the ACE2 receptor and the systemic inflammatory response elicited by the virus may contribute to renal injury (7).

The expression of the ACE2 receptor in healthy liver tissues suggests that liver injury may be due to direct viral invasion, leading to hepatocyte apoptosis. The role of endotheliitis in hepatic injury, where hypercoagulability can lead to thrombosis in the portohepatic system, resulting in liver tissue hypoxia, reactive oxygen species influx, and a proinflammatory state that contributes to liver damage. These insights suggest that liver injury in COVID-19 patients can be influenced by direct viral effects on liver tissues, leading to various pathological processes that contribute to liver dysfunction. Besides, the excessive immune response triggered by COVID-19 infection leads to an inflammatory cytokine storm, a critical factor in hepatic injury. Patients with COVID-19, especially those critically ill, commonly exhibit lower lymphocyte counts, higher levels of inflammatory cytokines, and specific risk factors for severe liver damage, such as increases in IL-6 and IL-10. The systemic inflammatory response syndrome in moderate to severe cases results in uncontrolled immune-mediated inflammation, contributing to secondary liver injury and potential multiorgan failure (8).

1.2 Multiorgan damage by sepsis

Sepsis is a syndrome of organ dysfunction that is critical for survival, which arises from an aberrant host response to a specific infection. The mortality rate associated with

doi:10.6342/NTU202402334

sepsis is 20–30% when multiple organ systems are impacted. Mortality in patients experiencing shock due to sepsis exceeds 40%. The manifestation of organ dysfunction resulting from this disrupted host response occurs when there is an increase in the sequential sepsis-related organ failure assessment (SOFA) score of 2 points or higher. The distribution of the indicated multiorgan dysfunction is not uniform. The cardiovascular and pulmonary systems are the most commonly impacted systems in approximately 50% of individuals. Approximately 30% of patients experience renal impairment (9).

Sepsis is commonly perceived as an exaggerated immune response to a pathogen, which triggers an intricate network of molecular cascades resulting in tissue destruction, organ failure, and, ultimately, mortality. The phenomenon encompasses both inflammatory and anti-inflammatory mechanisms, as well as humoral and cellular responses and impairments in the circulatory system (10). The immunological response results in impaired organ function, ultimately culminating in the development of multiple organ dysfunction syndrome (MODS) and mortality.

1.3 Mitochondrial damage causes multiorgan damage

Inflammatory cytokines are immunological reactions designed to eliminate infections. However, the excessive production of cytokines in a hyperinflammatory state can result in irreversible harm to cells and mitochondria, as well as the induction of cell death. This can potentially lead to further damage to organs (11). The emergence of inflammatory cytokine storms is a frequently reported phenomenon in the context of COVID-19 and sepsis. The primary locations of adenosine triphosphate (ATP) generation, known as mitochondria, play a crucial role in the regulation of cellular immunity, homeostasis, as well as cell survival and death (12). Research findings

doi:10.6342/NTU202402334

indicate that SARS-CoV-2 has the ability to hijack the mitochondria of immune cells, undergo replication within the mitochondrial framework, and disrupt mitochondrial dynamics, ultimately resulting in cell death (13). Mitochondria represent a significant focal point for SARS-CoV-2. On the other hand, since SARS-CoV-2 enters cells through angiotensin-converting enzyme 2 (ACE2), a crucial enzyme of the renin-angiotensin-aldosterone system (RAAS) that regulates blood pressure, electrolytes, and the inflammatory response, may also be a contributing factor to COVID-19-related organ damage (14).

During infection, SARS-CoV-2 induces an elevation in mitochondrial DNA (mtDNA) levels, potentially eliciting an exaggerated immune response and resulting in severe pathological manifestations of COVID-19, such as multiorgan failure (15). COVID-19 is also reportedly associated with inhibiting mitochondrial gene transcription, and patients infected with the virus have reported decreased bioenergetics and mitochondrial oxidative phosphorylation (OXPHOS). The infection caused by SARS-CoV-2 impedes the bioenergetic processes of mitochondria, potentially initiating inflammasome activation. As a result, the inhibition of mitochondria leads to an overproduction of cytokines and has a significant effect on organs that rely significantly on mitochondrial energy generation (16).

Further investigation is required to determine whether SARS-CoV-2 can affect organ function by direct viral infection via ACE2, mitochondrial damage, or multiorgan damage induced by an inflammatory cytokine storm. Multiple organ damage can also result from sepsis and other non-COVID-19 acute respiratory infections, and it is critical to comprehend how mitochondrial damage contributes to this mechanism.

doi:10.6342/NTU202402334

1.4 Long COVID in COVID-19 and non-COVID-19 acute respiratory infections

Long COVID is a highly heterogeneous disease that can occur after an individual has been infected with SARS-CoV-2. It is characterized by the onset of new or persistent symptoms that appear more than four weeks after the initial infection. The most common symptoms are fatigue, breathlessness, and cognitive impairment. However, there are various recognized symptoms, making it challenging to describe, diagnose, and treat the condition accurately.

Non-COVID-19 acute respiratory infections are defined as self-reports of a hospital diagnosis of influenza, seasonal coronavirus, and other virus-caused illnesses or other upper/lower respiratory infections not caused by SARS-CoV-2. It also includes self-reported symptoms of acute respiratory infections accompanied by a negative SARS-CoV-2 swab test (17).

Both non-COVID-19 acute respiratory infections and COVID-19 can lead to symptoms of long COVID.

1.5 Research flow

To further investigate the difference between multiorgan damage caused by sepsis and COVID-19, the phenomenon of long COVID, the transcriptome and machine learning analyses are also applied to sepsis and other non-COVID-19 acute respiratory infections sample data. According to the collected evidence, it is hypothesized that COVID-19 can potentially harm the heart, kidney, and liver by causing malfunction in the mitochondria and triggering downstream reactions, in addition to directly infecting these organs and causing cytokine storms. A comprehensive transcriptome analysis was

doi:10.6342/NTU202402334

performed to investigate the involvement of mitochondria in multiorgan damage associated with COVID-19 (Fig 1). The research indicates a reasonable inference that there might be a link between SARS-CoV-2 infection-induced mitochondrial damage and subsequent deterioration of heart, kidney, and liver function, leading to multiorgan damage (1). This study facilitates the comprehension of the pathogenesis of multiorgan problems induced by SARS-CoV-2 and the formulation of therapeutic protocols.

In summary, the study started by gathering data with gene expression, which is then analyzed to identify genes that show significant differences in expression. Next, toxicity analysis is conducted to pinpoint crucial genes that exhibit significant differential expression in connection with related factors. With the genes from significant differential expression and toxicity analysis as features, machine learning was implemented to validate the hypothesis. The sepsis sample set was used to compare sepsis and COVID-19 in multi-organ damage. And the non-COVID-19 acute respiratory infections sample set was used for the further investigation of the symptom of long COVID. Through the association with mitochondria, machine learning was used to validate the prediction that SARS-CoV-2 would cause substantial damage to cardiac, hepatic, and renal function. This was accomplished using the genes identified as significantly differentially expressed through statistical and biological approaches. A review of existing literature examined the common genes, the importance of features, and pathway analysis of these transcriptomes. The aim was to delve into the mechanisms through which SARS-CoV-2 affects mitochondria and might further harm the heart, liver, and kidney function. Conclusions were drawn based on the findings.

doi:10.6342/NTU202402334

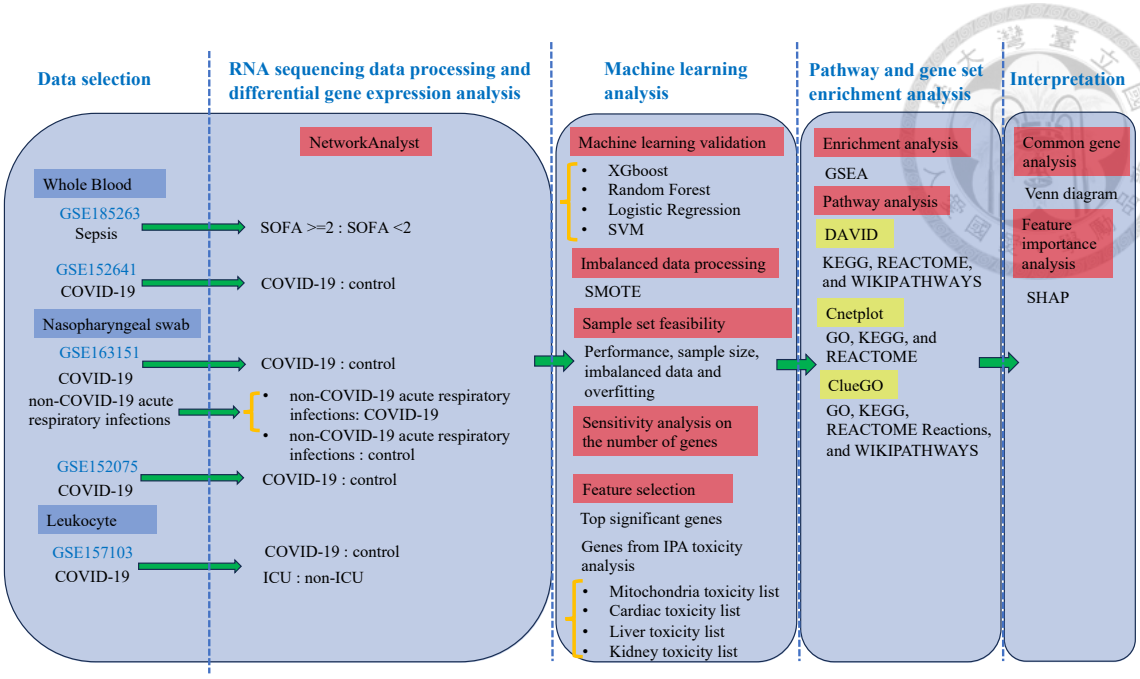


Fig 1. Flow chart of the research.

1.6 Specific aims of this research

This research has to achieve some specific aims.

Specific Aim 1 - Develop robust machine learning models to predict COVID-19 infection status based on the optimal data sample and significant gene sets :

The selection of sample sets in machine learning significantly impacts the outcomes and effectiveness of the resulting models. This procedure involves meticulously considering various factors such as machine learning performance, sample size, the status of imbalanced data, and overfit evaluation, ensuring that the sample set chosen is representative, unbiased, and capable of yielding accurate and generalizable predictions. The machine learning model's performance is inherently tied to the quality and appropriateness of the sample set. A well-curated sample set ensures the model can learn effectively and generalize well to unseen data. Performance metrics are often used to assess the model's performance on the chosen sample set.

doi:10.6342/NTU202402334

Sensitivity analysis plays a pivotal role in fine-tuning the machine learning process, particularly when determining the minimal set of significant genes required as a baseline for predicting COVID-19 infection status. Sensitivity analysis involves systematically varying the number of significant genes and observing the impact on the model's performance. This helps identify the optimal subset of genes that still provides a robust and reliable prediction, thereby optimizing the feature set for the machine learning model. By doing so, researchers can reduce the complexity of the model and enhance its interpretability and efficiency.

Specific Aim 2 - Investigate the multifaceted impact of COVID-19 and sepsis on vital organs using machine learning models trained on well-curated sample sets :

This approach further tests the association between COVID-19 and its effects on vital organs such as mitochondria, heart, kidneys, and liver. By leveraging the selected sample sets and significant genes, machine learning models can be trained to identify patterns and correlations that may not be evident through traditional analytical methods. This analysis is crucial in understanding the multifaceted impact of COVID-19 on different bodily systems.

Additionally, the mechanisms leading to multiorgan damage in both COVID-19 and sepsis are examined to investigate the differences in the pathological processes induced by these conditions. Sepsis shares some similarities with COVID-19 regarding its potential to cause widespread inflammation and organ damage. However, the specific pathways and molecular mechanisms involved may differ. By comparing and contrasting the effects of these two conditions using machine learning models trained on well-curated sample sets, we can better understand how each disease affects the body and identify potential targets for therapeutic intervention.

doi:10.6342/NTU202402334

These models can then explore the complex associations between COVID-19 and its effects on various organs, understand the mechanisms leading to multiorgan damage in COVID-19 and sepsis. Through these efforts, machine learning can provide powerful tools and insights for addressing the challenges posed by COVID-19 and improving patient outcomes.

Specific Aim 3 - Identify critical mitochondria-related genes and pathways associated with long COVID and multiorgan damage through pathway and toxicity analyses :

Through pathway and toxicity analyses of significantly differentially expressed genes from COVID-19 and other acute respiratory infections to identify key differences and similarities for long COVID. This comparative approach helps isolate the unique factors contributing to long COVID.

The toxicity analysis specifically examines genes related to mitochondrial function to understand how COVID-19 and sepsis impact mitochondrial health and contribute to multiorgan damage. The heart, liver, kidney, and mitochondria-related toxicities associated with COVID-19 identify lists of genes linked to these organ-specific toxicities and select them for further machine learning analysis based on their potential roles in COVID-19. This process involves integrating data on organ damage observed in sample sets and analyzing the underlying genetic factors contributing to this damage.

The common genes between the toxicity lists for the heart, liver, kidney, and mitochondria across these organ systems indicate shared pathways or mechanisms related to multiorgan damage in COVID-19. In addition to identifying common genes, the analysis assesses the importance of the ranked feature of significantly differentially expressed genes as biomarkers. This step involves evaluating how each gene contributes to the observed toxicities in the heart, liver, kidney, and mitochondria.

doi:10.6342/NTU202402334

1.7 Significance of the work

Similar to COVID-19 and sepsis, non-COVID-19 respiratory infections can induce cytokine storms, leading to multiorgan involvement and failure. Besides the well-known roles of ACE2 receptors and inflammatory cytokine storms, mitochondrial pathways are crucial in COVID-19 pathogenesis. Mitochondria are involved in energy production, cellular metabolism, and the regulation of apoptotic pathways. SARS-CoV-2 infection can disrupt mitochondrial function, leading to increased oxidative stress, impaired energy metabolism, and subsequent cellular and tissue damage. This mitochondrial dysfunction contributes significantly to the multiorgan impairments observed in COVID-19 patients, which concluded that long COVID is related to mitochondrial damage.

Machine learning facilitates the testing of statistical and biological hypotheses, thus verifying the complex interactions between various genetic and biological factors. By incorporating gene list adjustment, machine learning served as a verification tool to facilitate testing biological hypotheses. A subsequent comprehensive investigation was undertaken to analyze gene and pathway networks, examining the potential association between COVID-19 and deficits in cardiac, hepatic, and renal functions through mitochondrial infection. It validates and explores the complex interactions between various biological and genetic factors involved in COVID-19, sepsis, and other acute respiratory infections. By analyzing datasets, machine learning can identify correlations without traditional statistical methods. This facilitates testing hypotheses related to the pathways and mechanisms of multiorgan damage, helping to uncover potential biomarkers and therapeutic targets.

doi:10.6342/NTU202402334

The comprehensive study of multiorgan damage mechanisms in COVID-19, sepsis, and non-COVID-19 acute respiratory infections is essential for understanding long COVID and developing effective treatments. The study of long COVID necessitates a comprehensive understanding of the mechanisms of multiorgan damage caused by COVID-19 and other non-COVID-19 acute respiratory infections. Researchers can identify critical pathways and genetic factors in these conditions by employing machine learning to verify statistical and biological hypotheses. Mitochondrial pathways, in particular, play a significant role in the multiorgan impairments seen in COVID-19, highlighting the need for targeted therapeutic interventions.

doi:10.6342/NTU202402334

Chapter 2 Materials and methods

The bioinformatics and machine learning techniques were employed to analyze multiple publicly available RNA-Seq data from clinical samples. The results of the analysis have provided valuable insights into the mechanisms underlying the biological processes of interest, thus contributing to the advancement of knowledge in this field.

2.1 Bioinformatics and machine learning tools

To facilitate the investigation, several cutting-edge tools were availed, including NetworkAnalyst, DAVID, IPA, ClueGO and GSEA.

NetworkAnalyst is a comprehensive web-based tool designed to facilitate the process of visual analytics, network analysis, and meta-analysis of gene expression data in bioinformatics (18).

DAVID (The Database for Annotation, Visualization and Integrated Discovery) which is a web-based bioinformatics resource designed to interpret the biological significance of gene expression data. This platform integrates various bioinformatics tools, allowing researchers to perform enrichment analysis on gene lists, identify biological functions and pathways, and conduct gene ontology classification. The capabilities of DAVID are the Gene Ontology (GO) analysis, which examines the distribution of genes across biological processes, cellular components, and molecular functions. It also offers pathway analysis using databases like KEGG and REACTOME to explore how genes participate in specific metabolic or signaling pathways. Additionally, the platform includes the analysis of protein interactions and the association of genes with specific diseases (19, 20).

doi:10.6342/NTU202402334

IPA (Ingenuity Pathway Analysis) is an analysis and development tool for genomics, proteomics, drug toxicology, clinical trials, and metabolic and regulatory pathway studies. It is an essential biological and medical research tool for investigating complicated biological systems (21).

ClueGO is a plug-in for Cytoscape designed to enhance the analysis of large lists of genes or proteins by elucidating the biological themes underlying them. The core functionality of ClueGO lies in its ability to conduct comprehensive gene ontology and pathway analysis. It groups functionally related gene ontology terms and biological pathways, allowing researchers to identify the predominant biological functions and pathways associated with their gene lists. ClueGO integrates the latest updates from major ontology and pathway databases such as Gene Ontology, KEGG, REACTOME, and WikiPathways (22, 23).

GSEA (Gene Set Enrichment Analysis) is a computational method used to determine if a predefined set of genes shows statistically significant differences between two biological states, such as different phenotypes. This method is crucial in genomics, particularly in the analysis of microarray and RNA-Seq data, as it aims to uncover the molecular mechanisms that distinguish various biological conditions. Unlike traditional methods that examine individual gene expression changes, GSEA evaluates groups of genes, known as gene sets, which share common biological functions, chromosomal locations, or regulatory pathways. This approach allows it to identify subtle but biologically significant changes in gene expression that might be overlooked when analyzing genes individually.

GSEA ranks all the genes in a dataset based on their correlation with a specific phenotype and then calculates an enrichment score (ES). This score measures the degree of overrepresentation of a gene set at the top or bottom of the ranked list, with a higher

doi:10.6342/NTU202402334

absolute score indicating significant enrichment. To ensure robust analysis, GSEA includes normalization steps to account for variations in gene set sizes. This provides a normalized enrichment score (NES) that facilitates comparisons across gene sets. Additionally, the method incorporates corrections for multiple hypothesis testing, typically using false discovery rate (FDR) or family-wise error rate (FWER), to enhance the accuracy of significance assessments (24).

The cnetplot function is a regularly utilized feature of the clusterProfiler R package, mainly employed to visualize outcomes from gene set enrichment analysis. This tool generates a category network plot by combining a category enrichment plot with a gene network plot. The clusterProfiler version is 4.8.3, and it is compatible with R version 4.3.1. Machine learning with Python packages including XGBoost (1.7.5), imblearn (0.10.1), and sklearn (1.2.2) were run under Python 3.10.11. Venn diagrams and feature importance related tools to investigate specific genes, aiming to unveil the biological pathways and important molecules in the heart, kidney, and liver that are associated with mitochondria affected by SARS-CoV-2.

2.2 Data availability

From the NCBI Gene Expression Omnibus (GEO) database, the raw counts of the RNA-seq data were acquired. Several criteria influence the selection of sample sets, which include the ratio of experimental and control groups, the sample size for machine learning, the type and status of the disease, and the kind of human tissue under study. Additional sample sets were subjected to evaluation prior to the selection of these four sets. However, most of these datasets have a restricted number of samples, usually significantly less than 50. Therefore, following the assessment according to the

doi:10.6342/NTU202402334

established criteria, the chosen sample sets consist of GSE152075, GSE163151, GSE157103, and GSE152641, which were selected for the purpose of identifying genes that exhibit significant differential expression and conducting toxicity analysis. Another sample set GSE185263 was selected to compare COVID-19 and sepsis.

The criteria for selecting sample sets for machine learning include machine learning performance, sample size, imbalanced data state, and overfitting assessment. A comparative and comprehensive analysis was conducted for the machine learning on GSE152075, GSE163151, GSE157103, and GSE152641.

Sample sets GSE152075 and GSE163151 were obtained from nasopharyngeal swabs, whereas GSE157103 was obtained from leukocytes, and GSE152641 and GSE185263 was obtained from whole blood (Table 1). Subsequently, machine learning techniques were employed on GSE152075, a dataset with a bigger sample size, to substantiate the impact of COVID-19 on mitochondria and to ascertain further other impacts on the heart, kidney, and liver. Similarly, GSE185263 was utilized to validate the effect of sepsis on multiple organ damage, while GSE163151 employed the same process as GSE152075 to examine other non-COVID-19 acute respiratory infections.

Table 1. Summary of RNA sequence sample sets of sepsis, COVID-19 and other non-COVID-19 acute respiratory infections in Gene Expression Omnibus. Reuse and modified from ref. (1).

GEO accession	GSE title	Tissue	Platform	Sample size
GSE185263	Predicting sepsis severity at first clinical presentation: the role of endotypes and mechanistic signatures(25)	Whole Blood	GPL16791 Illumina HiSeq 2500 (Homo sapiens)	392 (Sepsis : control = 348 : 44) (SOFA >=2 : SOFA <2 = 207 : 138, SOFA NA = 47)

doi:10.6342/NTU202402334

GSE163151	A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood (26)	Nasopharyngeal swab	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)	269 (non-COVID-19 acute respiratory infections : influenza, seasonal coronavirus, and other viruses) (COVID-19 : control = 138: 11) (non-COVID-19 acute respiratory infections : control = 120 : 11)
GSE152075	In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age (27)	Nasopharyngeal swab	GPL18573 Illumina NextSeq 500 (Homo sapiens)	484 (COVID-19 : control = 430: 54)
GSE157103	Large-scale Multiomic Analysis of COVID-19 Severity (28)	Leukocyte	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)	126 (COVID-19 : control = 100: 26; ICU : non-ICU = 50: 50)
GSE152641	Transcriptomic Similarities and Differences in Host Response between SARS-CoV-2 and Other Viral Infection (29)	Whole blood	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)	86 (COVID-19 : control = 62: 24)

doi:10.6342/NTU202402334

2.3 RNA sequencing data processing and differential gene expression analysis

The RNA sequencing data is processed by NetworkAnalyst for gene expression analysis with the tuning of various parameters. In gene expression analysis, the mean is used for gene-level summarization. This simplifies the representation, reduces random noise, and makes the data more suitable for further analysis. Additionally, it helps address issues with multiple testing and assumes that each transcript contributes equally to the gene's activity.

Prior to conducting differential gene expression (DGE) analysis, filtering is used to increase statistical power by removing genes that do not show a response. To obtain reliable and meaningful results from the data, it is essential to use appropriate normalization techniques. Normalization is essential for effectively identifying transcriptional variations and maintaining uniform expression distributions across all samples in the experiment.

To establish the parameters for variance and abundance filters, filtering is needed to adjust the number of genes that are erroneous or unlikely to be informative and excluded from further analysis. The variance filter removes features that have a percentile rank of variance lower than the threshold, indicating that they have constant expression values across different contexts. In this case, the variance filtering was set to a threshold of 15, following the default setting of NetworkAnalyst. As a result, data that falls within the lowest 15th percentile of expression will be excluded. The criterion used to remove features with counts below the specified threshold is known as low

doi:10.6342/NTU202402334

abundance. The default value of 4 is used in NetworkAnalyst for this investigation. Log2-counts per million (Log2-CPM) is a frequently employed normalization technique in processing RNA-seq data. Log transformation aims to reduce the range of data so that genes with significant expression fluctuation do not have a disproportionate impact on subsequent analysis.

The Limma (Linear Models for Microarray Data) approach is frequently employed in the field of differential expression analysis. This approach is favored for its use of linear models, which offer enhanced computational efficiency in comparison to alternative methods for studying differential expression. By leveraging linear models, researchers can more effectively and accurately identify genes or other molecular entities whose expression levels vary across different experimental conditions (30). The threshold for the adjusted p-value was established at 0.05 (31), and the $\log_2$-fold change was adjusted to 1.5 (32) in order to increase the stringency of the analysis and to ensure that the genes identified as significantly differentially expressed possess both statistical significance and biological relevance with an appropriate level of variance.

2.4 Pathway and gene set enrichment analysis

Following the identification of significantly differentially expressed genes using NetworkAnalyst's differential gene expression analysis, a comparative analysis of the sample sets was conducted using Ingenuity Pathway Analysis (IPA; version 84978992). From this analysis, toxicity lists for each sample set were obtained using the tox analysis.

In addition to IPA tox analysis, pathway enrichment analysis, and functional annotation were performed using DAVID, Cnetplot, ClueGO, and GSEA. Using DAVID, the transcriptomes

doi:10.6342/NTU202402334

for genes with significant differential expression according to KEGG, REACTOME, and WIKIPATHWAYS were examined. Genes are associated in enriched pathways with GO, KEGG, and REACTOME through the use of the cnetplot function in the clusterProfiler package of R. Pathways involving GO terms, KEGG, REACTOME Reactions, and WIKIPATHWAYS were functionally annotated and analyzed in the study using ClueGO. Among the parameters are a p-value threshold of 0.05, a Gene Ontology (GO) hierarchy from 8 to 15, and a criterion for selecting pathways consisting of 2 genes at 6% each. In order to assess the impact of predefined gene sets on a phenotype, GSEA sorts gene tables according to correlation with the phenotype. The analysis was conducted using the h.all.v2023.1.Hs.symbols.gmt gene sets database. The database was used to collapse gene symbols in a total of 1000 permutations. The chip platform used was Human Gene Symbol with Remapping MSigDB.v2023.1.Hs.chip, and the permutation type was phenotypic. By default, the other parameters are not changed. To further investigate the biological meaning of the transcriptomes in the liver, kidneys, heart, and mitochondria, a Venn diagram (33) was employed to locate common genes. In addition, differentially expressed genes, either up or down-regulation, were shown using a volcano plot (34) by VolcaNoseR.

2.5 Imbalanced data processing

In machine learning analysis, imbalanced data is a significant problem that will lead to biased results. The imbalanced data issue might be caused by the small size of the control group in the GSE152075 sample set due to the 430:54 ratio of the experimental group to the control group. As a result, the minority samples were analyzed, and more samples were added to the set using SMOTE (Synthetic Minority Oversampling

doi:10.6342/NTU202402334

Technique). Oversampling minority classes is one example of an imbalanced sample problem that SMOTE attempts to solve. One class has a much larger number of samples than the other in an imbalanced sample set. In the pursuit of maximizing overall accuracy, many machine learning models end up favoring the majority class over those of the minority, either by failing to recognize or incorrectly label them. To solve the problem of class imbalance, SMOTE generates new synthetic samples instead of just replicating ones from the minority class. It identifies k-nearest neighbors, all members of the same minority class, for every sample in that class. The next step is to pick a neighbor randomly and then create a new sample somewhere between this neighbor and the synthetic sample. The technique is repeated until the minority class reaches the target sample count or proportion. Using the following formula, we can characterize the SMOTE synthesis method:

The k nearest neighbors of a minority class sample x_i, represented as x_{zi}, are randomly chosen. And then, select a random number λ from the range of 0 to 1. It is possible to produce the new synthetic sample x_{new} by applying the following formula as

$$x_{\text{new}} = x_i + \lambda \times (x_{zi} - x_i)$$

According to this procedure, the new synthetic sample will reside in the line segment between the original sample and its chosen neighbor. Results may vary with each iteration due to the randomized nature of λ.

When it comes to dealing with imbalanced data, SMOTE has significant benefits. To increase the variety of the sample set and decrease the risk of overfitting, synthetic samples are generated rather than duplicate data from minority classes. Improved prediction accuracy is another benefit of SMOTE's expansion of minority class data,

doi:10.6342/NTU202402334

allowing models better to comprehend the class's intricacies (35). The procedure used a 1:1 ratio between the experimental and control groups.

2.6 Machine learning algorithms

There are 4 machine learning algorithms employed in this research, including XGBoost, Random Forest, Logistic Regression, and SVM.

XGBoost (eXtreme Gradient Boosting) is a powerful and efficient machine learning algorithm that is widely used for supervised learning tasks. It implements the boosting method based on decision tree algorithms and offers outstanding capabilities for classification, regression, and ranking tasks. Some of the key features and advantages of XGBoost include its ability to handle large datasets with ease, its efficient handling of missing data, its feature for regularization to prevent overfitting, its support for parallel computation which speeds up the training process.

XGBoost builds upon the foundation of gradient boosting to minimize the disparities between predicted and actual values. In addition to standard gradient boosting features, XGBoost incorporates L1 and L2 regularization to prevent overfitting and enhance generalization. XGBoost can handle missing values automatically and exhibits excellent scalability for data size and model complexity, making it suitable for a wide range of applications from small to large datasets. It can also perform cross-validation during each boosting round to continuously monitor model performance and prevent overfitting (36). This study set the parameters for colsample_bytree=0.9, learning_rate=0.1, max_depth=10, n_estimators=50.

The Random Forest technique is widely used for classification and regression tasks. It works by creating multiple decision trees to form a "forest," with each tree

doi:10.6342/NTU202402334

independently learning from different subsets of the data during the training process. For classification, the Random Forest uses majority voting to determine the final category, while for regression, it averages the predictions from all trees to produce the final result, improving accuracy and model stability. Through bootstrap aggregating, also known as bagging, Random Forest performs multiple samplings of the original training dataset, using each sample to train a separate decision tree. This enhances the robustness of the model and helps prevent overfitting. The Random Subspace Method increases model diversity and reduces the correlation between trees by randomly selecting features during the construction of each tree.

Random Forest offers several advantages, including its efficient handling of high-dimensional data and tolerance to missing data. It can process large datasets with a lot of variables without needing feature reduction, making it a practical choice. Moreover, it maintains good predictive performance even when a significant portion of data is missing, further highlighting its robustness (37). This study set the parameters for max_depth=10, min_samples_split=5, n_estimators=100.

Logistic Regression is a method designed for solving binary classification problems, predicting one of two possible outcomes, such as yes or no. The aim of Logistic Regression is to model the probability that the target variable belongs to a specific category. This probability, which falls between 0 and 1, is calculated using the Logistic Regression, also known as the Sigmoid function. The Sigmoid function is an S-shaped curve that maps any real-valued number into a value between 0 and 1 but doesn't reach these extremes. The function is defined as

$$f(x) = \frac{1}{1+e^{-x}},$$

, where x is the value input into the function.

doi:10.6342/NTU202402334

When using Logistic Regression for predictions, a threshold is typically set to make a binary decision, most commonly at 0.5. If the model's predicted probability is greater than 0.5, the outcome is classified as category 1. If it is less than or equal to 0.5, it is classified as category 0. This method not only provides decision support but also explains clearly how each variable influences the predicted outcome. This is a significant advantage of using Logistic Regression as a classification tool (38). This study set the parameters for C=50, max_iter=5000.

SVM (Support Vector Machine) stands as a powerful and versatile machine learning algorithm that finds broad application in both classification and regression tasks. At its core, SVM aims to pinpoint a hyperplane that effectively separates a dataset into two distinct classes. The operational principles of SVM are anchored in several key concepts. Firstly, SVM endeavors to maximize the margin, which refers to finding a hyperplane that maximizes the distance from the nearest data points of each class, termed support vectors. This margin represents the minimum distance between the hyperplane and the support vectors. By maximizing this margin, SVM seeks to bolster the model's generalization capabilities and mitigate the risk of overfitting.

In higher-dimensional spaces, the concept of a hyperplane becomes fundamental. In two-dimensional data, this hyperplane could be a line; in three-dimensional data, it is a plane; and in higher dimensions, it is referred to simply as a hyperplane. SVM is particularly effective in handling datasets with numerous features, as it performs robustly in these high-dimensional spaces. Another critical feature of SVM is the kernel trick. This technique allows the SVM to operate in a higher-dimensional feature space without explicitly calculating the coordinates of the data within that space. By utilizing a kernel function to compute the dot product of data pairs in this feature space, SVM can effectively manage non-linear relationships in the original space. Common kernel

doi:10.6342/NTU202402334

functions include the linear, polynomial, and radial basis function (RBF) or Gaussian kernel.

In order to handle datasets that are not linearly separable, SVM incorporates the concept of a soft margin, allowing for a certain degree of misclassification. This involves setting a regularization parameter, typically denoted as C, to balance the trade-off between minimizing errors on the training data and maintaining a small margin. (39) This study set the parameters for kernel='rbf', C=100, gamma=0.01, probability=True.

To conduct validation and provide predictions for COVID-19. The K-fold cross-validation technique was employed, using 10-fold sets.

2.7 The indices of machine learning performance

Multiple indices, including accuracy, sensitivity, specificity, precision, F1-Score, MCC, and AUC, were assessed to measure the effectiveness of the models. Each index has its distinct formula, relevance, and application, especially in imbalanced data situations.

Accuracy is a key metric used to assess the performance of classification models. It measures the proportion of correct predictions, including true positives and true negatives, out of all cases tested. This metric shows how often the model predicts outcomes correctly, clearly indicating the model's overall effectiveness. The formula for calculating accuracy is straightforward, which is

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

However, accuracy has its limitations, particularly in datasets with highly imbalanced class distribution. For instance, in a dataset where one class dominates, a

doi:10.6342/NTU202402334

model could achieve high accuracy by simply predicting every instance as the majority class, thereby neglecting its predictive power on the minority class. In such scenarios, it's crucial to turn to more detailed performance metrics like precision, recall, and the F1-score. These metrics provide deeper insights into the model's performance across different classes, enlightening us about the model's true capabilities.

Despite these limitations, accuracy is still widely used in scenarios where the classes are well-balanced, and the costs of false positives and false negatives are roughly equal. It provides a convenient way to quickly assess a model's overall effectiveness in making correct classifications (40).

In the field of machine learning, sensitivity is a crucial metric used to evaluate the performance of classification models, especially in situations where identifying positive cases is vital, such as in medical diagnostics. Sensitivity, also known as the true positive rate or recall, primarily measures a model's ability to correctly identify actual positive cases.

Sensitivity is defined as the proportion of true positives to the total number of actual positives, which is the sum of true positives and false negatives. The formula for sensitivity is expressed as

$$Sensitivity = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})}$$

The significance of sensitivity is paramount in applications where capturing as many positive cases as possible is essential. In medical diagnostics, high sensitivity ensures the accurate identification of individuals with the condition, reducing the risk of missed diagnoses. However, sensitivity has limitations as it may increase the false positive rate, resulting in incorrect labeling of actual negatives as positives, potentially leading to unnecessary treatments or further testing in some cases. Therefore, when evaluating

doi:10.6342/NTU202402334

sensitivity, it is often considered alongside specificity, which measures the proportion of true negatives correctly identified to comprehensively assess the model's performance. In fields like healthcare and fraud detection, where missing a positive prediction is costly, sensitivity is a crucial metric. In summary, sensitivity is a valuable measurement tool that effectively reflects a model's capability in identifying positive cases, particularly in applications where accurate identification of positives is crucial (41).

Specificity is a very important metric used to evaluate the performance of classification models in machine learning, especially when accurately identifying negative cases is crucial. It is also known as the true negative rate and measures the proportion of actual negatives the model correctly identifies. This is particularly important in scenarios where avoiding false positives is critical, such as medical diagnostics, where a false positive result could lead to unnecessary or harmful treatments.

The formula for calculating specificity as

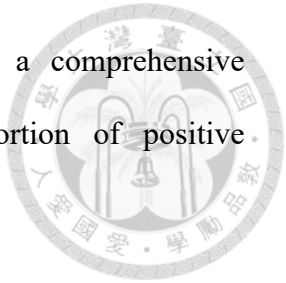
$$Specificity = \frac{\text{True Negatives}}{(\text{True Negatives} + \text{False Positives})}$$

This calculation reflects the model's ability to dismiss non-events or negatives correctly.

In a nutshell, specificity is a key metric for evaluating a classification model's ability to identify negative cases, especially when the cost of false positives is high. It is instrumental in adjusting the threshold settings of classification algorithms to strike the best balance between capturing true positives and avoiding false negatives, thereby significantly improving the model's diagnostic capability (41).

Precision is a crucial metric in machine learning used to measure the accuracy of classification models, particularly in situations where the cost of false positives is

doi:10.6342/NTU202402334



significant. In conjunction with recall (sensitivity), it provides a comprehensive evaluation of a model's performance by measuring the proportion of positive identifications made by the model that are actually correct.

The formula for precision is

$$Precision = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}$$

The significance of precision lies in its role in minimizing the rate of false positives. In medical diagnostics, achieving high precision is crucial to ensure that patients are not subjected to unnecessary treatments based on incorrect diagnoses.

However, focusing solely on precision may result in a trade-off with recall, as increasing precision could decrease recall, potentially causing some actual positive cases to be missed. High precision ensures that a model does not mislabel negative instances as positive, but it may also mean missing out on some positive cases.

In practice, precision is extensively used to fine-tune classification models, especially in decision-making processes where the consequence of a false positive is more detrimental than that of a false negative. It is instrumental in setting thresholds for classification decisions and in adjusting models to achieve desired levels of false positive rates (40).

The F1-Score is an essential metric in machine learning that is used to evaluate the performance of classification models, mainly when there are uneven class distributions or differences in the costs of false positives and false negatives. The harmonic mean of precision and recall provides a balanced measure that accounts for both aspects of a model's performance. And the F1-Score is calculated using the formula as

$$F1 - Score = 2\text{x}\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

doi:10.6342/NTU202402334

A higher F1-Score, approaching 1, suggests superior model performance by effectively balancing precision and recall, which is particularly critical in datasets with imbalanced classes. A low F1-Score, close to 0, suggests inadequate model performance. This will encourage models to maintain a balance between the two metrics.

Adjusting a model to achieve an ideal F1-Score can ensure that the model does not miss significant events or generate too many irrelevant alerts. In summary, the F1-Score is a comprehensive metric that reflects a model's ability to accurately identify true positives while avoiding false positives, especially in applications with low tolerance for misclassification. By balancing precision and recall, the F1-Score helps model developers understand and enhance the overall performance of their models (40).

The MCC (Matthews Correlation Coefficient) in machine learning is a crucial metric for assessing the performance of binary classification models. It is particularly effective in scenarios with imbalanced class distributions or when there are significant disparities in the costs associated with false positives and false negatives. The MCC ranges from -1 to +1, where +1 indicates perfect prediction, 0 suggests a prediction no better than random chance, and -1 indicates total disagreement between prediction and observation. This coefficient provides a comprehensive assessment by considering all four elements of the confusion matrix.

The MCC is calculated using the formula as

$$MCC = \frac{(\mathrm{TP\ x\ TN} - \mathrm{FP\ x\ FN})}{\sqrt{((\mathrm{TP} + \mathrm{FP})\ \mathrm{x}\ (\mathrm{TP} + \mathrm{FN})\ \mathrm{x}\ (\mathrm{TN} + \mathrm{FP})\ \mathrm{x}\ (\mathrm{TN} + \mathrm{FN})}}$$

Where TP, TN, FP, and FN are four elements of the confusion matrix represent true positives, true negatives, false positives, and false negatives, respectively.

The MCC is particularly crucial for datasets with class imbalance, where more straightforward metrics such as accuracy might be misleading. MCC offers a balanced

doi:10.6342/NTU202402334

measure of accuracy and reliability, making it suitable for comprehensive performance evaluations. In summary, the Matthews Correlation Coefficient is particularly valuable in addressing challenging classification issues, ensuring that models achieve high accuracy and balance predicting across different categories (42).

AUC (Area Under the Curve) in machine learning is an important measure used for evaluating the performance of classification models, especially in situations where class distributions are imbalanced. The AUC refers explicitly to the area under the Receiver Operating Characteristic (ROC) curve, which is a graph showing a binary classifier's diagnostic ability as its discrimination threshold is changed.

The ROC curve plots the true positive rate (TPR , also known as sensitivity or recall) against the false positive rate (FPR , 1-specificity) at different threshold settings. TPR measures the model's ability to correctly identify actual positives, while FPR measures the proportion of actual negatives that are incorrectly identified as positives. The AUC ranges from 0 to 1, with 0.5 indicating the model has no discrimination ability and 1.0 indicating perfect prediction. The AUC considers the entire area under the ROC curve, providing a comprehensive performance measure across all possible classification thresholds.

AUC is beneficial for comparing different models as it offers a single metric to describe overall performance across all threshold levels, thus helping to identify the best model for effectively separating the classes. Due to its reduced sensitivity to class imbalance, it is a reliable metric for assessing models on datasets with imbalanced classes (43).

These metrics are essential when the data is imbalanced, as they provide a more comprehensive understanding of a model's performance compared to just accuracy. While a high accuracy can be deceptive in these situations, a high score in F1, MCC,

doi:10.6342/NTU202402334

and AUC demonstrates that the model well manages both minority and majority classes, offering a comprehensive evaluation of its prediction ability across various class distributions.

2.8 Sample sets feasibility analysis for machine learning

Utilizing diverse sample sets can yield varying outcomes, introducing the research intricacy. The sample set GSE157103 was derived from leukocytes, while GSE185263 and GSE152641 were derived from whole blood samples. Similar to GSE152075, GSE163151 also utilized nasopharyngeal swabs as the primary source of samples. However, the substantial difference in the number of samples between the COVID-19 and control groups (138:11) in GSE163151 poses a challenge. Even employing techniques such as SMOTE to address the issue of imbalanced data may produce inadequate outcomes.

When training models with limited data, there is a greater likelihood of encountering a phenomenon known as overfitting. Overfitting is when a model performs well on the training data but fails to generalize effectively to unfamiliar data. This indicates that the model may have gotten excessively intricate and inflexibly assimilated the characteristics of the training data, resulting in poor performance when applied to unfamiliar data.

It is commonly observed in various machine learning tasks that the model's performance typically enhances with increased samples. This phenomenon can be linked to the model being given a greater quantity of data, allowing it to acquire a more extensive range of features and patterns. Increasing the amount of data usually leads to improved generalization capabilities of the model.

doi:10.6342/NTU202402334

The machine learning comparison was based on GSE152641, GSE152075, GSE157103, and GSE163151 sample sets, using the top 40 significantly differentially expressed genes. Considering the above factors, GSE152075 exhibited the most optimum selection regarding machine learning performance, sample size, imbalanced data state, and overfitting assessment (Table 2).

Table 2. The comparison of sample sets for machine learning.

Overall Ranking (5) = (1) +(2) + (3) + (4). Reuse and modified from ref. (1).

Machine Learning \ Feature		GSE152641	GSE152075	GSE157103	GSE163151
XGBoost	Accuracy	0.884	0.969	0.897	0.973
	Sensitivity	0.919	0.984	0.920	0.986
	Specificity	0.792	0.852	0.808	0.818
	Precision	0.919	0.981	0.948	0.986
	F1-Score	0.919	0.983	0.934	0.986
	MCC	0.711	0.842	0.700	0.804
	AUC	0.958	0.972	0.886	0.991
	(1) Performance Ranking	(4)	(1)	(3)	(1)
	(2) Sample Size Ranking	86 (4)	484 (1)	126 (3)	149 (2)
	(3) Imbalanced	(62: 24 => 27.91%) (1)	(430: 54 => 11.16%) (3)	(100: 26 => 20.63%) (2)	(138: 11 => 7.38%) (4)

doi:10.6342/NTU202402334

	Data Ranking				
	(4) Overfitting Ranking	Overfitting (2)	Pass (1)	Overfitting (2)	Pass (1)
	(5) Overall Ranking	(11) **4**	(6) **1**	(10) **3**	(8) **2**
Random Forest	Accuracy	0.930	0.969	0.865	0.980
	Sensitivity	0.952	0.991	0.890	1.000
	Specificity	0.875	0.796	0.769	0.727
	Precision	0.952	0.975	0.937	0.979
	F1-Score	0.952	0.983	0.913	0.989
	MCC	0.827	0.837	0.619	0.844
	AUC	0.995	0.961	0.939	0.917
	(1) Performance Ranking	(3)	(1)	(4)	(1)
	(2) Sample Size Ranking	86 (4)	484 (1)	126 (3)	149 (2)
	(3) Imbalanced Data Ranking	(62: 24 => 27.91%) (1)	(430: 54 => 11.16%) (3)	(100: 26 => 20.63%) (2)	(138: 11 => 7.38%) (4)
	(4) Overfitting Ranking	Overfitting (2)	Overfitting (2)	Overfitting (2)	Overfitting (2)
	(5) Overall	(10) **3**	(7) **1**	(11) **4**	(9) **2**

doi:10.6342/NTU202402334

	Ranking				
Logistic Regression	Accuracy	0.977	0.899	0.857	0.973
	Sensitivity	0.968	0.902	0.860	0.986
	Specificity	1.000	0.870	0.846	0.818
	Precision	1.000	0.982	0.956	0.986
	F1-Score	0.984	0.941	0.905	0.986
	MCC	0.945	0.628	0.633	0.804
	AUC	0.987	0.976	0.866	0.984
	(1) Performance Ranking	(3)	(2)	(4)	(1)
	(2) Sample Size Ranking	86 (4)	484 (1)	126 (3)	149 (2)
	(3) Imbalanced Data Ranking	(62: 24 => 27.91%) (1)	(430: 54 => 11.16%) (3)	(100: 26 => 20.63%) (2)	(138: 11 => 7.38%) (4)
	(4) Overfitting Ranking	Overfitting (2)	Pass (1)	Overfitting (2)	Pass (1)
	(5) Overall Ranking	(10) **3**	(7) **1**	(11) **4**	(8) **2**
SVM	Accuracy	0.953	0.800	0.865	0.973
	Sensitivity	0.952	0.786	0.860	0.993
	Specificity	0.958	0.907	0.885	0.727
	Precision	0.983	0.985	0.966	0.979
	F1-Score	0.967	0.875	0.910	0.986

doi:10.6342/NTU202402334

	MCC	0.889	0.480	0.662	0.790
	AUC	0.987	0.962	0.949	0.914
	(1) Performance Ranking	(3)	(2)	(4)	(1)
	(2) Sample Size Ranking	86 (4)	484 (1)	126 (3)	149 (2)
	(3) Imbalanced Data Ranking	(62: 24 => 27.91%) (1)	(430: 54 => 11.16%) (3)	(100: 26 => 20.63%) (2)	(138: 11 => 7.38%) (4)
	(4) Overfitting Ranking	Overfitting (2)	Pass (1)	Overfitting (2)	Overfitting (2)
	(5) Overall Ranking	(10) **3**	(7) **1**	(11) **4**	(9) **2**
Selection Result			**X**		

In addition, employing more sample sets can provide a more comprehensive perspective. However, conducting a detailed analysis of a single sample set enables a more nuanced study and comprehension. Hence, the primary utilization of the GSE152075 sample set, comprising 484 samples, was to ensure consistency and uniformity in the forthcoming study.

The machine learning algorithms utilized the selected genes from the GSE152075 sample set as features to examine if the relationship between COVID-19 and its impact on the mitochondria, heart, kidney, and liver could be predicted. The machine learning

doi:10.6342/NTU202402334

results were used further to analyze the pathways and biological implications of the genes.

2.9 Feature ranking and SHAP

In machine learning, feature ranking and SHAP (SHapley Additive exPlanations) are essential concepts that improve model interpretability and help researchers understand the impact of different features on model predictions.

Feature ranking is a technique used to determine the contribution of each feature in a dataset to the predictive performance of a machine learning model. This technique involves ranking features based on their effectiveness in improving model accuracy, aiding in feature selection, reducing model complexity, and enhancing model transparency. Methods for feature ranking include statistical correlation tests, tree-based methods (such as Random Forests and Gradient Boosting Machines), and mutual information scores. These methods help researchers focus on the most impactful variables, optimizing data collection and processing efforts.

SHAP is based on the concept of Shapley values from cooperative game theory, designed initially to fairly distribute the payoff of a game among its players based on their contributions. In machine learning, SHAP values explain how each feature contributes to the prediction made by the model for each instance, indicating how a feature positively or negatively impacts the target prediction. The advantages of SHAP include providing local explanations specific to each prediction and global interpretations, offering insights into the model's overall behavior. Additionally, the SHAP method ensures fairness and consistency in the contributions assigned to features, aligning them with their actual impact on model predictions.

doi:10.6342/NTU202402334

By enhancing understanding of how models operate, feature ranking and SHAP increase the credibility and practicality of machine learning models. Feature ranking guides the initial development of models, ensuring they focus on the most relevant features. SHAP, meanwhile, helps further validate and refine models, ensuring that outputs meet expectations and effectively explaining the reasons behind these outputs. (44).

SHapley Additive exPlanations, 0.41.0 was used to analyze the prediction interpretation of the contribution of each feature.

doi:10.6342/NTU202402334

Chapter 3 Results

Upon conducting an analysis and processing of the raw count data from RNA sequencing for sample sets GSE185263, GSE163151, GSE152075, GSE157103, and GSE152641, the genes of significance for each sample set were identified and subsequently subjected to machine learning, pathway analysis, and IPA-Tox analysis. The core of the analysis is the relationship between multiorgan damage and mitochondria dysfunction in COVID-19. In contrast, sepsis can also result in multiorgan damage.

The sample data of sepsis was analyzed for comparison in order to comprehend the mechanism and distinctions between multiorgan damage caused by COVID-19 and sepsis. And the purpose of the analysis of the sample data of non-COVID-19 acute respiratory infections, which include influenza, seasonal coronavirus, and other viruses, is to identify the correlation between mitochondria and COVID-19 in long COVID-19.

3.1 Sensitivity analysis of machine learning in COVID-19 and sepsis samples

The first step in the analysis was to evaluate the predictive capabilities of the features identified by the statistical methodology regarding sepsis, COVID-19, and other non-COVID-19 acute respiratory infections in order to gain insight into the prospective utility of machine learning. As a result, the adjusted p-values of the significantly differentially expressed genes were ranked from the differential gene expression analysis of GSE152075 in NetworkAnalyst. Machine learning and pathway analysis were conducted using the top 100 significantly differentially expressed genes from the

doi:10.6342/NTU202402334

differential expression analysis of GSE152075, which were sorted according to the adjusted p-value for each gene (Table 3).

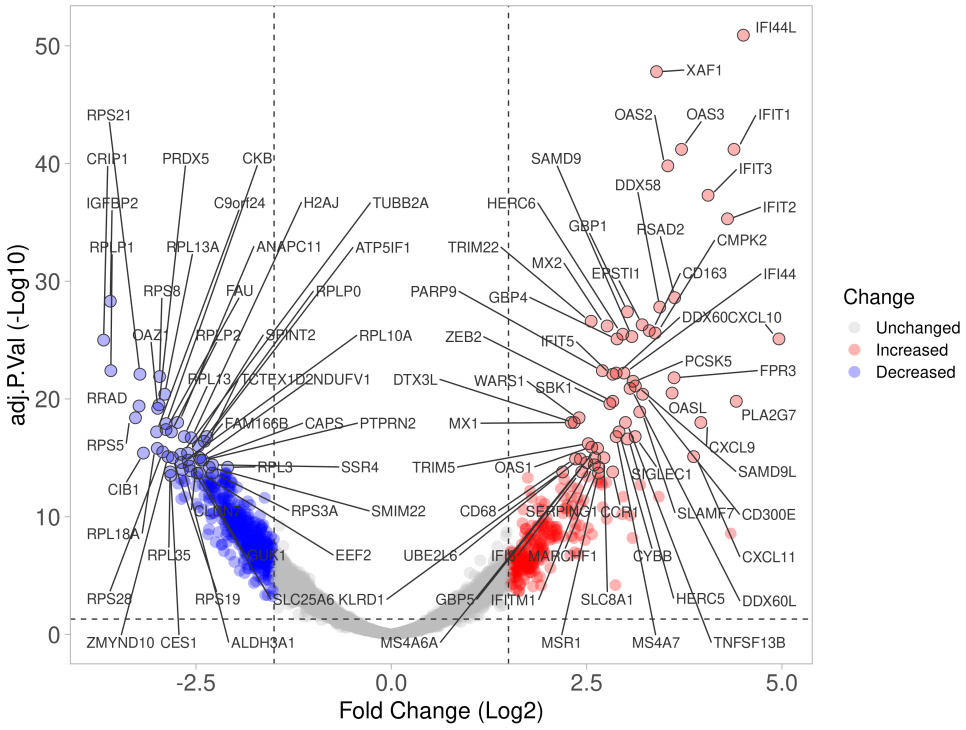
Table 3. Top 100 significantly differentially expressed genes from GSE152075.

Reuse and modified from ref. (1).

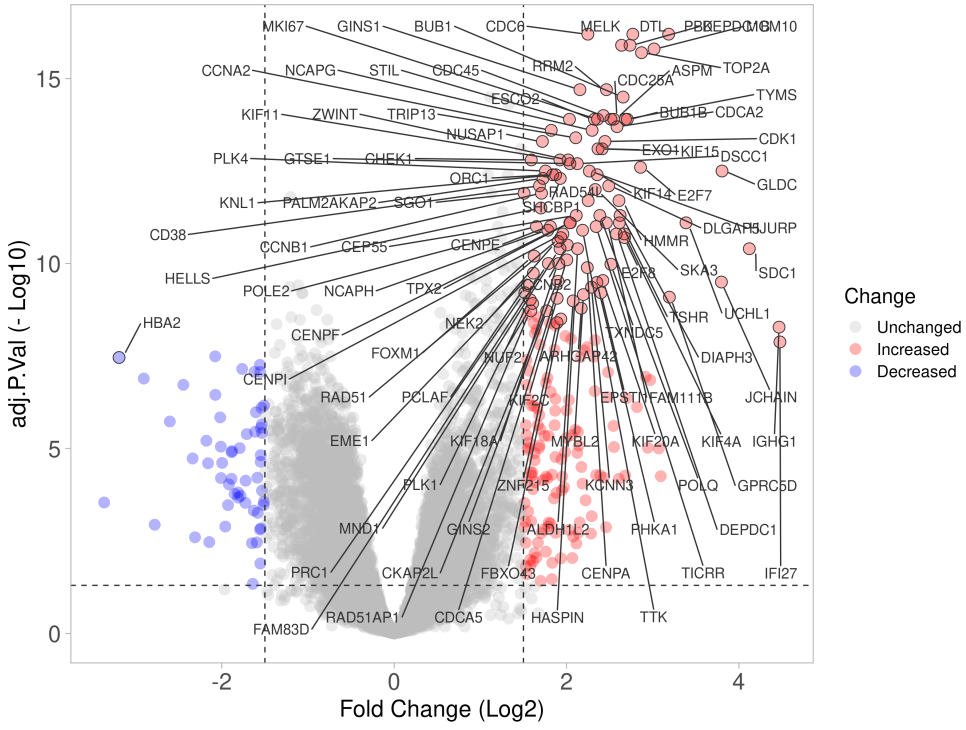
IFI44L, XAF1, IFIT1, OAS3, OAS2, IFIT3, IFIT2, RSAD2, IGFBP2, DDX58,
GBP1, TRIM22, EPSTI1, MX2, CD163, CMPK2, HERC6, SAMD9, CXCL10, GBP4,
CRIP1, PARP9, RPLP1, DDX60, IFI44, IFIT5, RPS21, RPS8, FPR3, PCSK5,
SAMD9L, DDX60L, OASL, RPL13A, CD300E, PLA2G7, ZEB2, SBK1, PRDX5, RRAD,
OAZ1, SLAMF7, RPS5, WARS1, ANAPC11, CXCL9, MX1, TNFSF13B, DTX3L, CKB,
FAU, CYBB, RPLP2, C9orf24, ATP5IF1, RPL13, SIGLEC1, MS4A7, H2AJ, HERC5,
TCTEX1D2, TRIM5, RPLP0, OAS1, RPL18A, MS4A6A, RPS28, SPINT2, CIB1, TUBB2A,
ZMYND10, CXCL11, NDUFV1, SLC8A1, SERPING1, RPS19, CD68, UBE2L6, CAPS, PTPRN2,
FAM166B, GBP5, RPL10A, IFITM1, SSR4, SLC25A6, SMIM22, MARCHF1, RPL3, ALDH3A1,
CLDN7, RPL35, KLRD1, CCR1, IFI6, MSR1, EEF2, RPS3A, GUK1, LAMTOR4

A volcano plot was illustrated to show down-regulation and up-regulation differential expression genes (Fig 2a, 2b, 2c, 2d).

doi:10.6342/NTU202402334

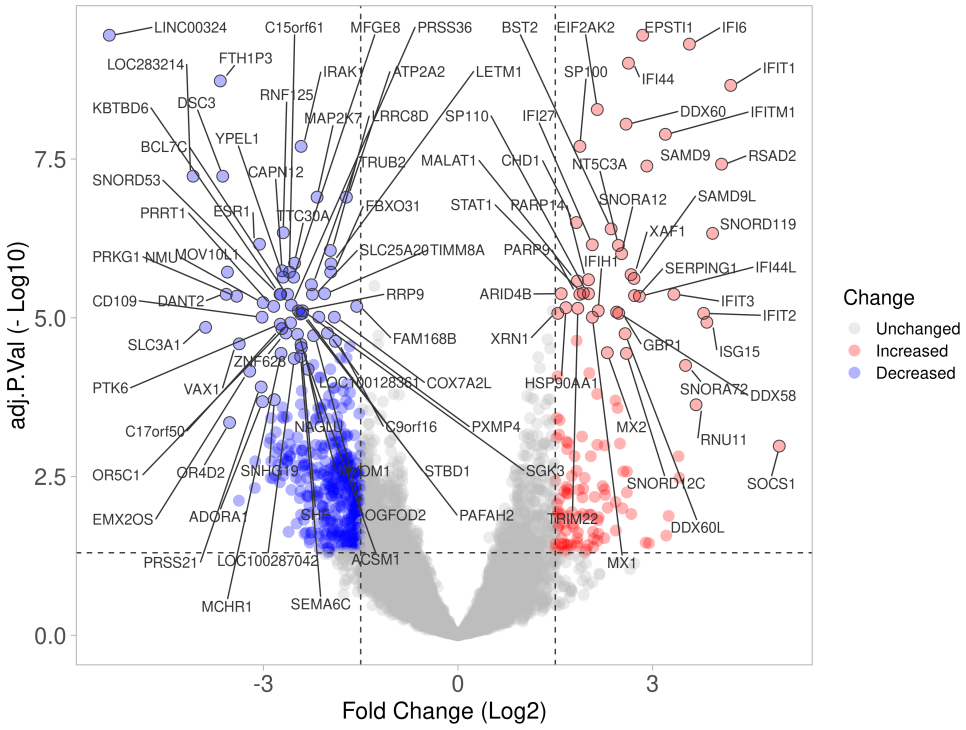


a.

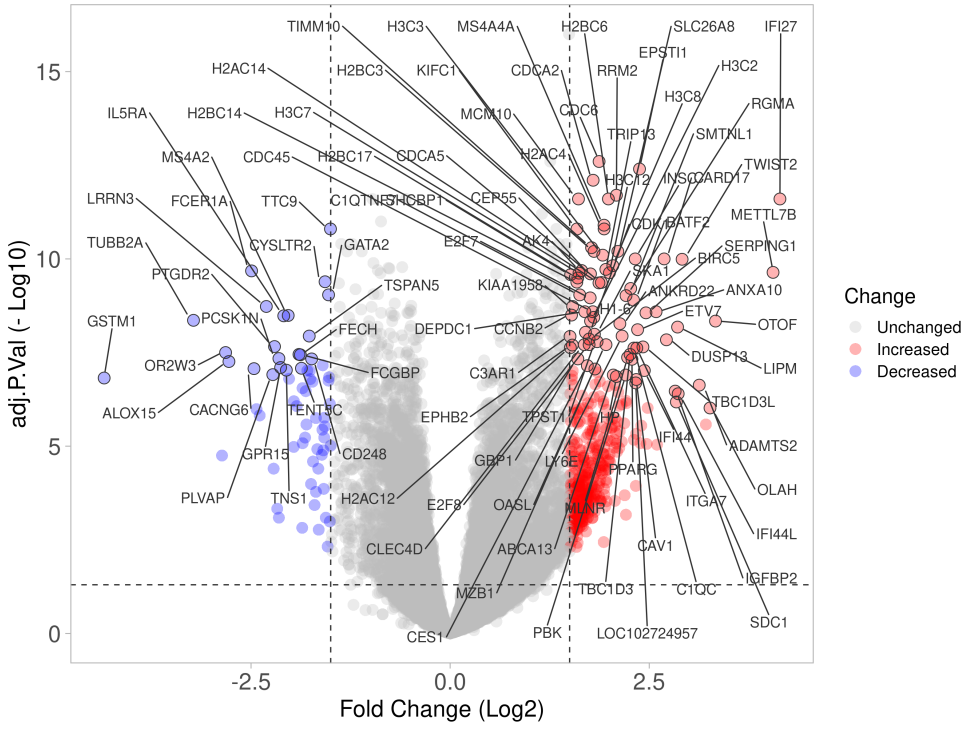


b.

doi:10.6342/NTU202402334



c.



d.

Fig 2. Volcano plot of up- and down-regulated differential expression genes.

Volcano plot (a) GSE152075, (b) GSE157103, (c) GSE163151, and (d) GSE152641 by VolcaNoseR. Reuse and modified from ref. (1).

doi:10.6342/NTU202402334

The machine learning prediction power for predicting COVID-19 was tested on the top 100, 60, 40, 30, and 20 significantly differentially expressed genes. The accuracy, F1-Score, and AUC of the four machine learning algorithms, except for SVM, were near or above 90% for the top 30–100 significantly differentially expressed genes. This suggests that 30 significantly differentially expressed genes are sufficient for good prediction power (Table 4). Nevertheless, the overall performances of 40 genes significantly differentially expressed in these four machine learning algorithms are superior to those of 30 genes significantly differentially expressed. Consequently, 40 genes were further investigated for a more thorough pathway enrichment analysis.

To verify the predictive capabilities of the features identified by the statistical methodology for COVID-19, a baseline of 40 randomly selected genes from all genes of GSE152075 were also tested as machine learning features. The findings indicated that the predictive capabilities of machine learning cannot be determined without the selection of preprocessed features (Table 4).

Research has demonstrated that the severity of sepsis in patients with early sepsis can be predicted by five distinct endotypes: Neutrophilic-Suppressive/ NPS, Inflammatory/ INF, Innate-Host-Defense/ IHD, Interferon/ IFN, and Adaptive/ ADA. Endotypes refer to specific groupings of diseases distinguished by their unique pathobiological mechanisms and differential gene expression signatures. The gene expression signatures associated with a poor prognosis Neutrophilic-Suppressive/NPS and Inflammatory (INF) endotypes are particularly useful in predicting the severity of sepsis (25). Furthermore, the presence of sepsis endotypes in COVID-19 patients would provide more evidence that indicates the severe COVID-19 disease is similar to sepsis. Sepsis can cause extensive tissue damage, organ malfunction, and, ultimately, multiple organ failures due to an exaggerated immune response. The primary factors contributing to organ failure in

doi:10.6342/NTU202402334

sepsis include the inflammatory response, microvascular dysfunction, blood clot formation, immunological suppression, inter-organ impacts, etc. The Sepsis 3 definitions characterize sepsis as a dysregulated immune response to infection and place greater emphasis on measuring organ failure in sepsis. The Sepsis-3 guidelines utilize the Sequential Organ Failure Assessment (SOFA) score to quantify the severity of the disease and as a tool to stratify the fatality risk. The SOFA score assesses the functioning of six organ systems, assigning values ranging from 0 (indicating no failure) to 4 (indicating severe dysfunction). A SOFA score of 2 or above, in the presence of infection and organ failure, provides strong evidence for diagnosing sepsis (45). And the top 40 significantly differentially expressed genes of GSE185263 were derived to distinguish the status of organ failure in patients based on the SOFA score >=2 definition.

The severity of sepsis may serve as a predictive indicator for the severity of COVID-19 (46). However, the status of COVID-19 vs health control seems not to be the predictive indicator for the severity of sepsis, or it might be because of the tissue specific issue (Table 4).

Table 4. Sensitivity analysis of machine learning with different samples and counts of genes.

The top/randomly selected significantly differentially expressed genes from GSE152075 and GSE185263 with training on GSE152075 and GSE185263 sample data. Reuse and modified from ref. (1).

doi:10.6342/NTU202402334

Machine Learning \ Features		100 genes from GSE152075 with GSE152075	60 genes from GSE152075 with GSE152075	40 genes from GSE152075 with GSE152075	30 genes from GSE152075 with GSE152075	20 genes from GSE152075 with GSE152075	Randomly selected 40 genes from GSE152075 with GSE152075	40 genes from GSE185263 with GSE152075	40 genes from GSE152075 with GSE185263
XGBoost	Accuracy	0.963	0.965	0.969	0.961	0.950	0.851	0.919	0.614
	Sensitivity	0.979	0.984	0.984	0.977	0.967	0.921	0.977	0.710
	Specificity	0.833	0.815	0.852	0.833	0.815	0.296	0.463	0.471
	Precision	0.979	0.977	0.981	0.979	0.977	0.912	0.935	0.668
	F1-Score	0.979	0.980	0.983	0.978	0.972	0.917	0.956	0.689
	MCC	0.812	0.819	0.842	0.804	0.758	0.225	0.534	0.185
	AUC	0.973	0.978	0.972	0.970	0.943	0.657	0.814	0.616
Random Forest	Accuracy	0.965	0.969	0.969	0.965	0.952	0.849	0.928	0.577
	Sensitivity	0.988	0.991	0.991	0.988	0.984	0.919	0.986	0.662
	Specificity	0.778	0.796	0.796	0.778	0.704	0.296	0.463	0.449
	Precision	0.973	0.975	0.975	0.973	0.964	0.912	0.936	0.643
	F1-Score	0.980	0.983	0.983	0.980	0.974	0.915	0.960	0.652
	MCC	0.815	0.837	0.837	0.815	0.745	0.220	0.577	0.112
	AUC	0.975	0.963	0.961	0.964	0.956	0.646	0.865	0.654
Logistic Regression	Accuracy	0.938	0.913	0.899	0.897	0.798	0.698	0.862	0.594
	Sensitivity	0.944	0.916	0.902	0.898	0.781	0.726	0.888	0.560
	Specificity	0.889	0.889	0.870	0.889	0.926	0.481	0.648	0.645
	Precision	0.985	0.985	0.982	0.985	0.988	0.918	0.953	0.703
	F1-Score	0.964	0.949	0.941	0.939	0.873	0.810	0.919	0.624
	MCC	0.737	0.669	0.628	0.631	0.487	0.143	0.448	0.201
	AUC	0.976	0.973	0.976	0.962	0.943	0.691	0.785	0.589
SVM	Accuracy	0.880	0.829	0.800	0.725	0.643	0.446	0.864	0.528
	Sensitivity	0.867	0.821	0.786	0.700	0.607	0.426	0.909	0.348
	Specificity	0.981	0.889	0.907	0.926	0.926	0.611	0.500	0.797
	Precision	0.997	0.983	0.985	0.987	0.985	0.897	0.935	0.720
	F1-Score	0.928	0.895	0.875	0.819	0.751	0.577	0.922	0.469
	MCC	0.638	0.511	0.480	0.408	0.337	0.023	0.376	0.156
	AUC	0.975	0.968	0.962	0.957	0.962	0.654	0.868	0.671

doi:10.6342/NTU202402334

3.2 Tissue specific issue in samples

The analytical process included performing preliminary analysis utilizing transcriptomics analysis, applying machine learning to evaluate the results obtained from the analysis results, and cross-comparing reports. This process effectively identified and validated the hypotheses. It is crucial to note that the validation results may be influenced by several factors related to the sample data, including the types of tissues examined, the detection platforms employed, and the sizes of the samples. The machine learning models developed using the top 40 significant genes from GSE152075 were evaluated by other sample sets, including GSE163151, GSE157103, and GSE152641. The prediction accuracy is significantly low, indicating that the trained machine learning model cannot be used in other sample sets if there are differences in tissues and gene expression detection platforms (Table 5).

Table 5. The comparison of learned machine learning models in different sample data.

GSE152075 with top 40 significant genes test in GSE163151, GSE157103, and GSE152641 sample data. Reuse and modified from ref. (1).

Features / Prediction accuracy of learned ML models from GSE152075	Top 40 significant genes of GSE152075 with GSE163151	Top 40 significant genes of GSE152075 with GSE157103	Top 40 significant genes of GSE152075 with GSE152641

doi:10.6342/NTU202402334

XGBoost	0.315	0.754	0.721
Random Forest	0.215	0.619	0.581
Logistic Regression	0.765	0.556	0.453
SVM	0.832	0.563	0.383

3.3 Pathway analysis in different sample sets

The top 40 significantly differentially expressed genes (Table 6) in sample set GSE152075/GSE163151/ GSE157103/GSE152641 for COVID-19 vs health control, GSE185263 for SOFA score >=2 vs SOFA score <2 and GSE163151 for non-COVID-19 acute respiratory infections vs health control/COVID-19 were analyzed for further pathway analysis.

Table 6. Top 40 significantly differentially expressed genes.

From GSE185263, GSE152075, GSE163151, GSE157103, and GSE152641. * Significantly differentially expressed genes, fold change +-1.5, p-value 0.5; GSE185263 sepsis fold change +-0.5, p-value 0.5. Reuse and modified from ref. (1).

GSE185263 SOFA score >=2 vs SOFA score <2	**GSE152075 COVID-19 vs health control**	**GSE163151 COVID-19 vs health control**	**GSE163151 non-COVID-19 acute respiratory infections vs COVID-19**	**GSE163151 non-COVID-19 acute respiratory infections vs health control**	**GSE157103 COVID-19 vs health control**	**GSE152641 COVID-19 vs health control**
33 significantly expressed genes; 5 Down	841 significantly expressed genes; 552 Down	574 significantly expressed genes; 438 Down	188 significantly expressed genes; 131 Down 57 Up	1435 significantly expressed genes; 1274 Down 161 Up	273 significantly expressed genes; 53 Down	491 significantly expressed genes; 65 Down

doi:10.6342/NTU202402334

28 Up*	289 Up	136 Up			220 Up	426 Up
MAFG, CKAP4, GADD45A, RETN, SLC51A, MMP9, QSOX1, CYP19A1, TRPM2, SPATC1, TP53I3, ARHGEF17, SEMA6B, ITGAM, MCEMP1, ARG1, HK3, NOS1AP, PPARG, RPS6KA5, NELL2, ACVR1B, HLA-DMB, KLF14, MICOS10, WIPI1, HP, TNFAIP8L3, CD40LG, GPR84, HLA-DQA1, TIMP4, RGL4, PGD, KMT5A, PCOLCE2,	*IFI44L, XAF1, IFIT1, OAS3, OAS2, IFIT3, IFIT2, RSAD2, IGFBP2, DDX58, GBP1, TRIM22, EPSTI1, MX2, CD163, CMPK2, HERC6, SAMD9, CXCL10, GBP4, CRIP1, PARP9, RPLP1, DDX60, IFI44, IFIT5, RPS21, RPS8, FPR3, PCSK5, SAMD9L, DDX60L, OASL, RPL13A, CD300E, PLA2G7, ZEB2, SBK1, PRDX5, RRAD*	*EPSTI1, LINC00324, IFI6, IFI44, FTH1P3, IFIT1, EIF2AK2, DDX60, IFITM1, IRAK1, SP100, RSAD2, SAMD9, LOC283214, DSC3, ATP2A2, MAP2K7, CHD1, BST2, RNF125, SNORD119, ESR1, IFI27, NT5C3A, LRRC8D, SNORA12, TTC30A, LETM1, CAPN12, MOV10L1, FBXO31, C15orf61, XAF1, YPEL1, MFGE8, SAMD9L,*	*SNORD87, SNORD12C, SNORD99, SNORD76, SNORD61, SNORD8, SNORD12B, SNORD55, SNORD47, SCARNA11, SNORD58A, NBPF25P, SNORD60, SNORD81, RNY5, SNORD105B, SNORD41, SNORD29, ADH1C, SNORA73A, SNORA24, SNORD68, SNORD110, SNORD79, SNORD12, SNORD95, BATF3, SNORD26, SNORD48, ADH7, SNORD46, SNORA44, SNORD27, PFN2, PLD4,*	*LINC00324, DANT2, RSAD2, SNORD110, SNORD83B, ZNF628, IFIT3, RNY5, SNORD32A, SNORD48, IFIT2, SLC3A1, OR10A4, CXCL10, DSC3, SNORD35A, ESAM, GBP5, EYA2, ISG15, GBP1, SHF, ATP2A2, MAP2K7, LRRC8D, IRAK1, SNORA10, SNORD58A, SNORD59B, SNORD71, SCUBE3, FTH1P3, LOC646903, SNORD105B, IFITM1, COX7A2L, PRRT1, LINC01011, GLIS3, SLF NL1*	*CDC6, PBK, DTL, DEPDC1B, MELK, MCM10, TOP2A, RRM2, GINS1, MCM6, BUB1, CLSPN, CDC45, ESCO2, CDC25A, ASPM, MKI67, STIL, TYMS, BUB1B, CDCA2, CCNA2, NCAPG, TRIP13, NUSAP1, CDK1, MCM4, KIF15, EXO1, ZWINT, CHEK1, KIF11, GTSE1, DSCC1, E2F7,*	*POLQ, CDC6, CLSPN, EPSTI1, CDCA2, RRM2, MCM10, H2BC6, IFI27, DHCR24, MS4A4A, H2AC4, DTL, TTC9, KIFC1, P2RY10, APOBEC3A, NT5E, CC2D2A, TIMM10, TOP1MT, GTSE1, SLC26A8, AMIGO1, GPR141, H2BC3, H3C3, INSC, CARD17, TWIST2, H2BC9, H3C8, TMEM144, TRIP13, AK4, FCER1A, HPD,*

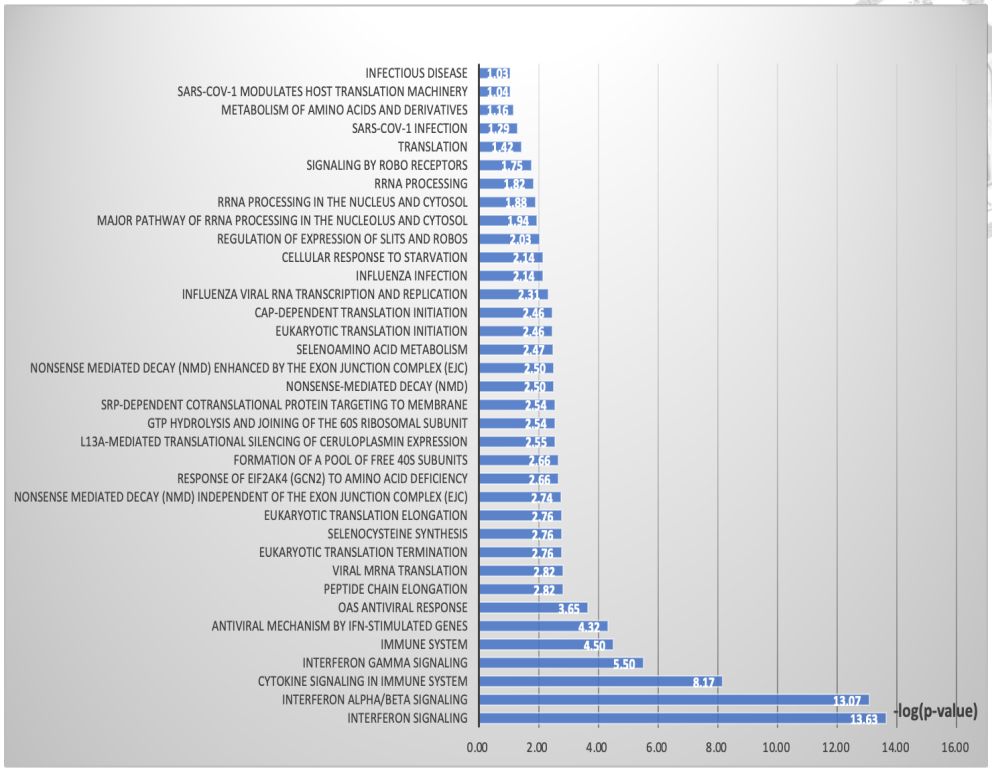
doi:10.6342/NTU202402334

GYG1, C1orf226, ADAMTS3, PFKFB2		*PARP14, MALAT1, TRUB2, AP3B1*	*SNORD21, ALDH1A1, CYP2F1, SNORD14D, SNORA14B*		*HJURP, PLK4, GLDC, NDC80, ORC1*	*METTL7B, H3C12, H3C7*

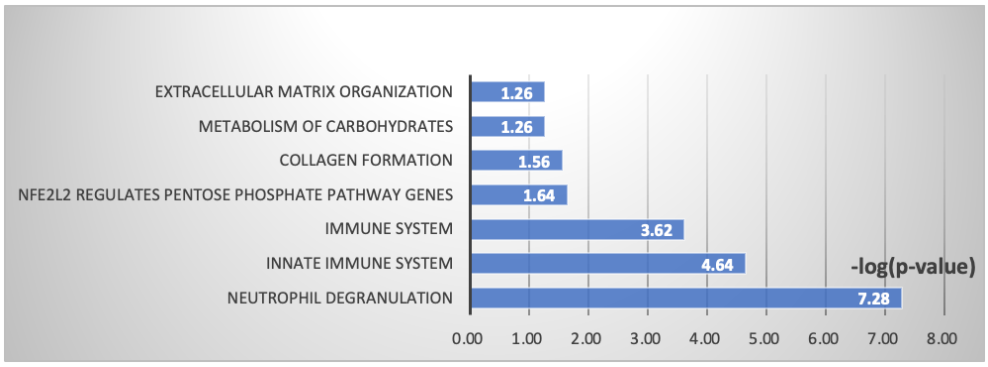
In DAVID REACTOME pathway analysis of the top 40 significantly differentially expressed genes revealed pathways associated with interferon signaling (47), interferon alpha/beta/gamma signaling, and cytokine signaling in the immune system, were identified. Furthermore, SARS-CoV-related pathways were identified as some of the major pathways in the dataset of GSE152075 for COVID-19 vs health control (Fig 3a). The pathway analysis in GSE185263 for SOFA score >=2 vs SOFA score <2, SOFA score >=2 is in the presence of organ failure for solid evidence for diagnosing sepsis, showed neutrophil degranulation and immune system were the major pathways for organ failure in sepsis (Fig 3b). In the dataset of GSE163151 for non-COVID-19 acute respiratory infections vs health control, the pathway analysis is similar to COVID-19 vs health control, which were interferon signaling, interferon alpha/beta/gamma signaling, and cytokine signaling in the immune system as the dominant pathway (Fig 3c). On the other hand, in the dataset of GSE163151 for non-COVID-19 acute respiratory infections vs COVID-19, the pathway analysis is mainly oxidation and metabolism. Mitochondria are the main energy conversion organs in cells, and one of their core functions is to generate energy through oxidation-reduction reactions. In the significantly differentially expressed genes of non-COVID-19 acute respiratory infections vs COVID-19, the main result of pathway analysis is oxidation related pathway, which shows mitochondria play an important role in COVID-19 (Fig 3d).

doi:10.6342/NTU202402334

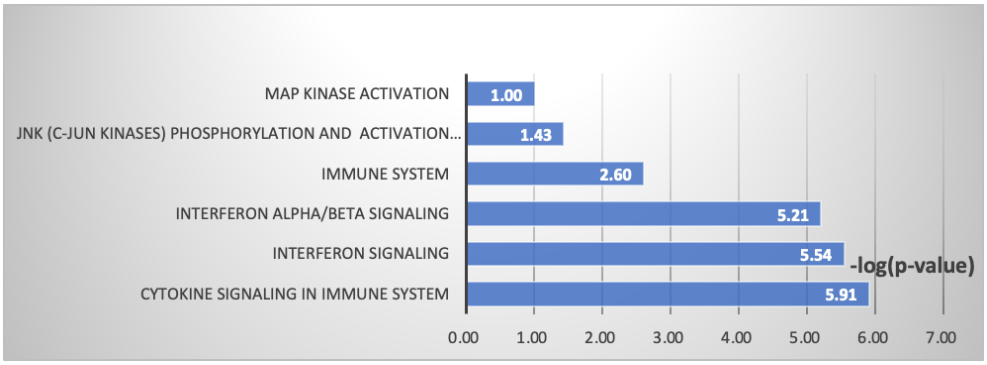
a.



b.



c.



doi:10.6342/NTU202402334

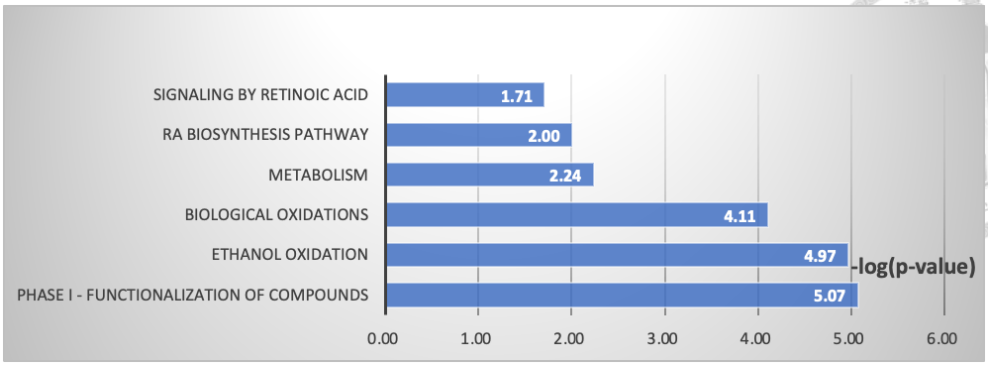
d. 

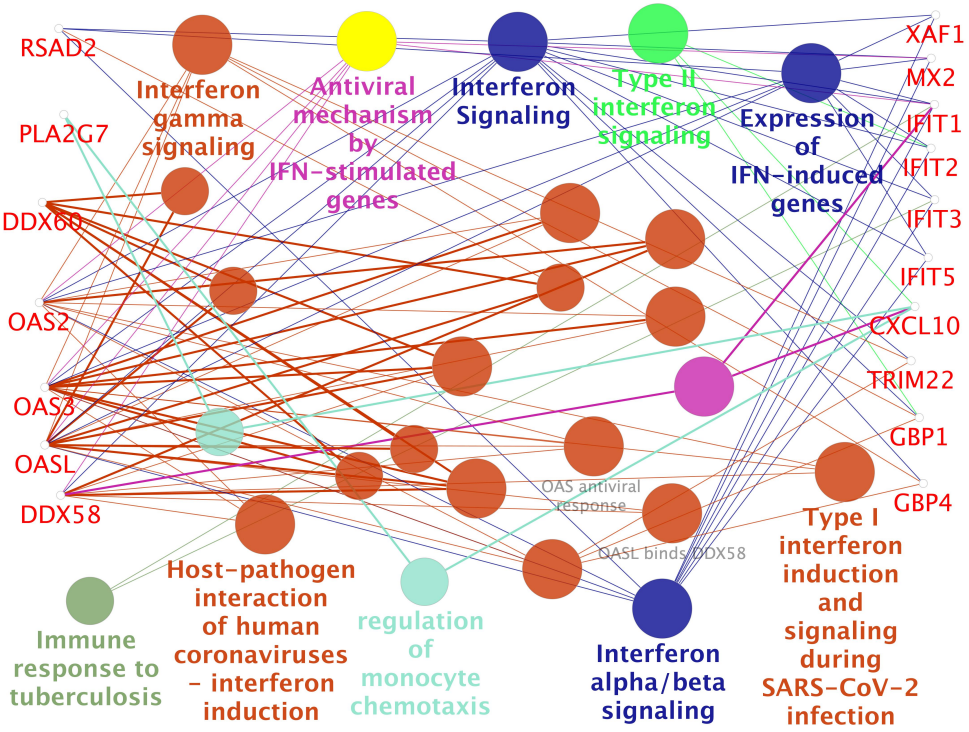
Fig 3. DAVID REACTOME pathway analysis of the top 40 significantly differentially expressed genes.

(a) for COVID-19 vs health control in GSE152075, (b) GSE185263 for SOFA score >=2 vs SOFA score <2, (c) GSE163151 for non-COVID-19 acute respiratory infections vs health control, and (d) non-COVID-19 acute respiratory infections vs COVID-19. Reuse and modified from ref. (1).

Further pathway analysis is conducted with ClueGo pathway network analysis integrated with GO, KEGG, WIKIPATHWAYS, REACTOME reactions/pathways and Cnetplot of enrichment analysis in Biological Process, KEGG, and REACTOME for the top 40 significantly differentially expressed genes. Interferon signaling, interferon alpha/beta/gamma signaling, immune response and SARS-CoV-2-related pathways were still the main pathway of GSE152075 for COVID-19 vs health control (Fig 4a, 4b, 4c, 4d). In GSE185263 for SOFA score >=2 vs SOFA score <2, hormone secretion and interferon-related pathways, neutrophil degranulation were the major pathways for organ failure in sepsis (Fig 4e, 4f, 4g, 4h). Interferon, response to virus and immune response were the main pathways in GSE163151 for non-COVID-19 acute respiratory infections vs health control (Fig 4i, 4j, 4k, 4l). On the other hand, GSE163151 for non-COVID-19 acute respiratory infections vs COVID-19, oxidation and metabolism related

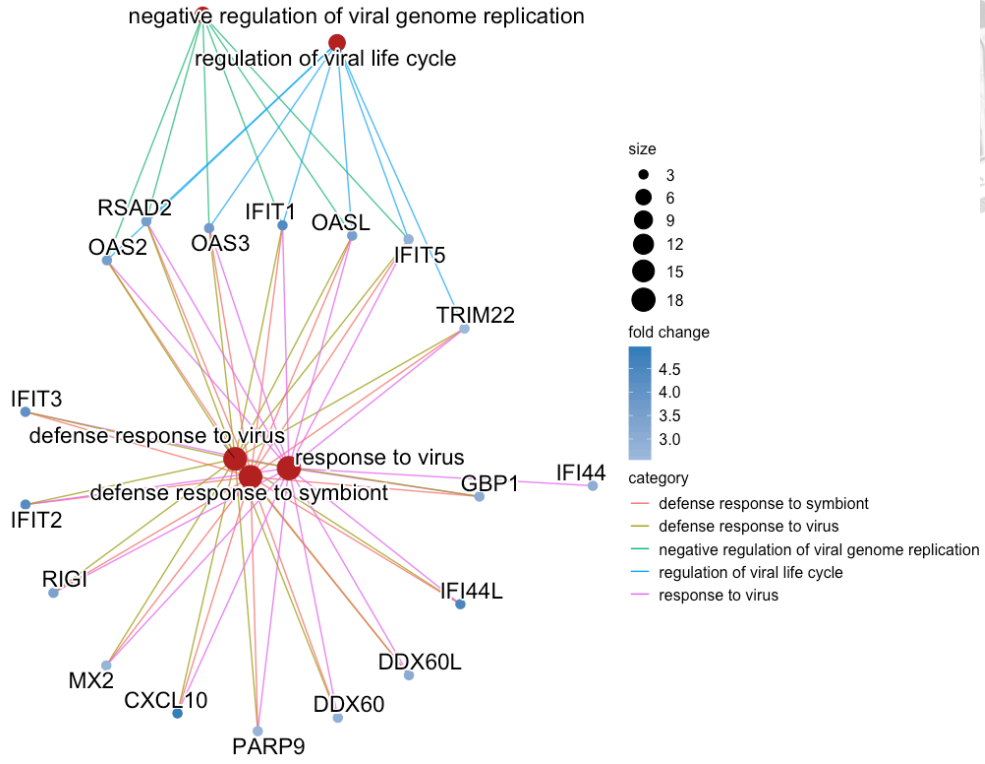
doi:10.6342/NTU202402334

pathways are the main pathways for the result (Fig 4m, 4n, 4o, 4p). And the top 40 significantly expressed genes in GSE152075 with GSEA also showed interferon gamma/alpha signaling are the main pathways (Fig 4q, 4s).

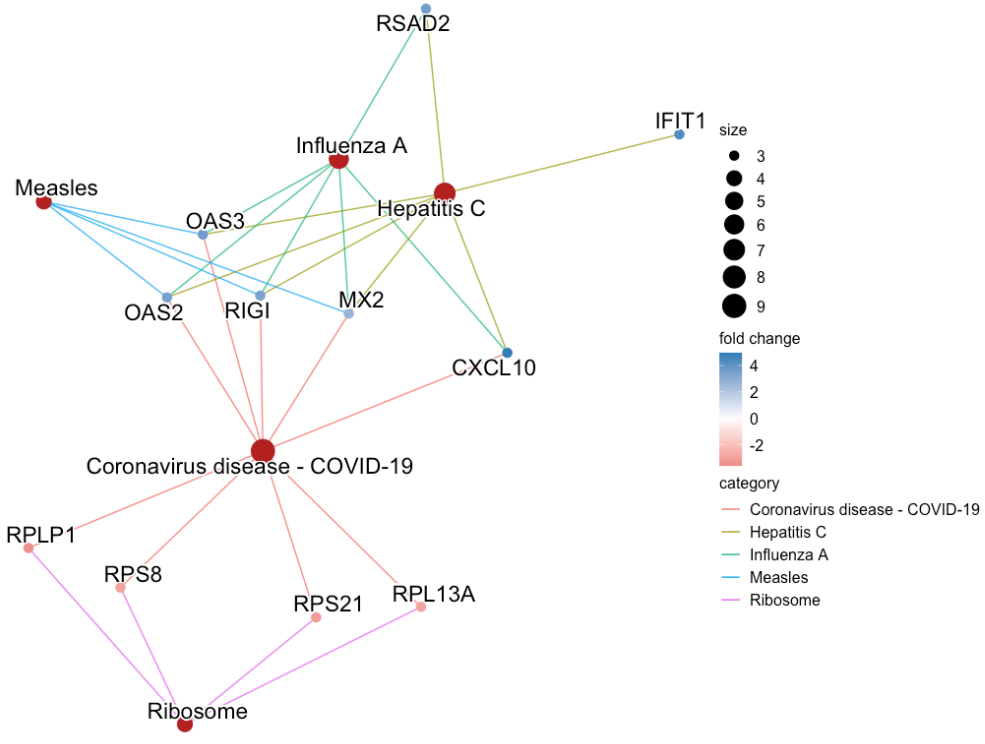


a.

doi:10.6342/NTU202402334

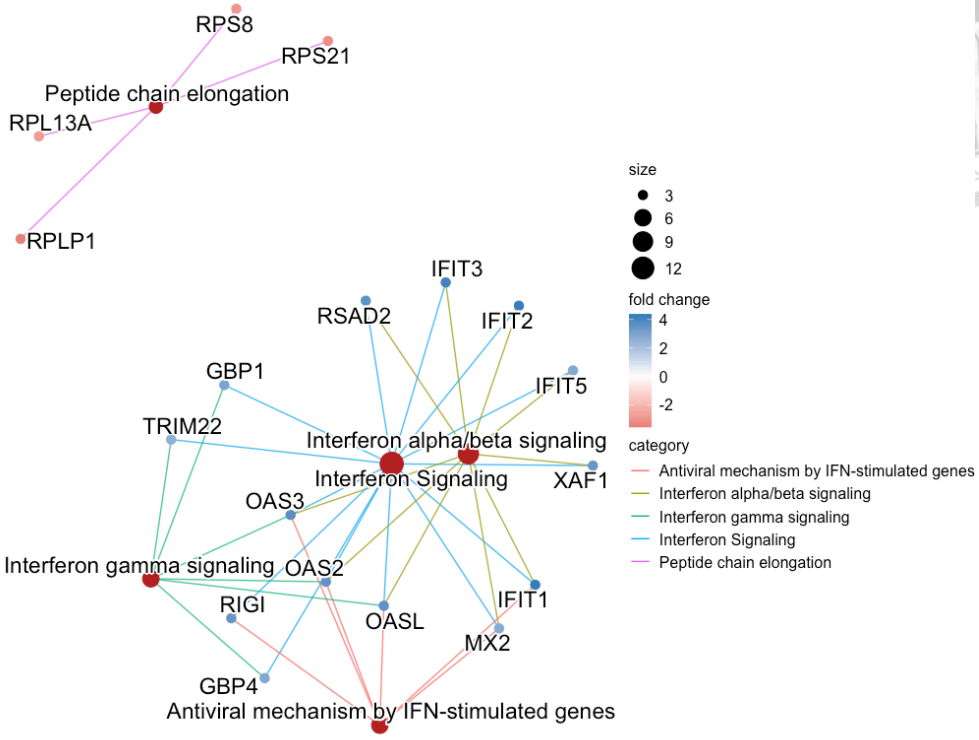


b.

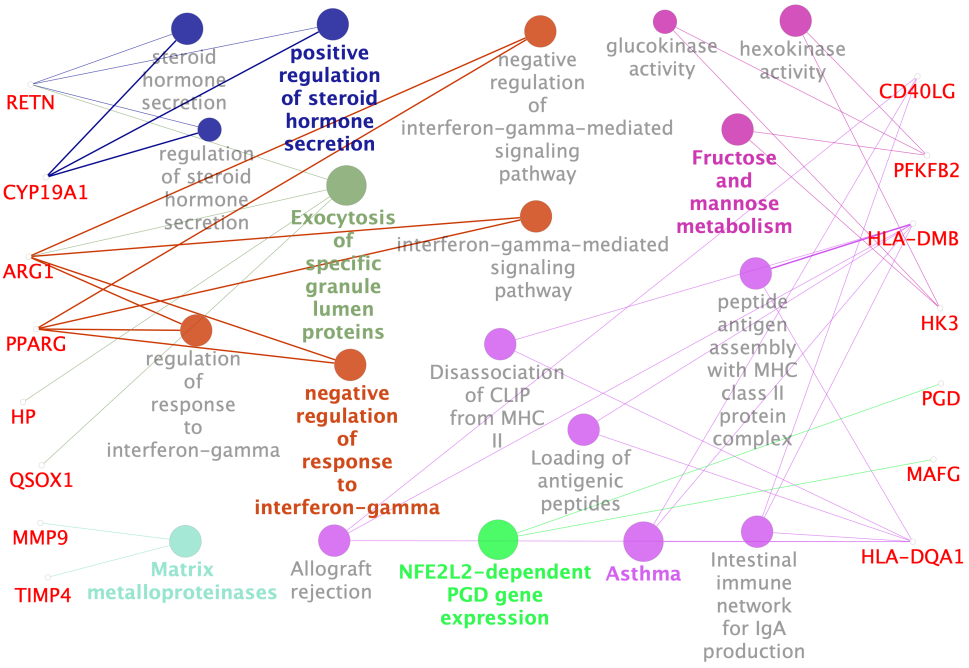


c.

doi:10.6342/NTU202402334

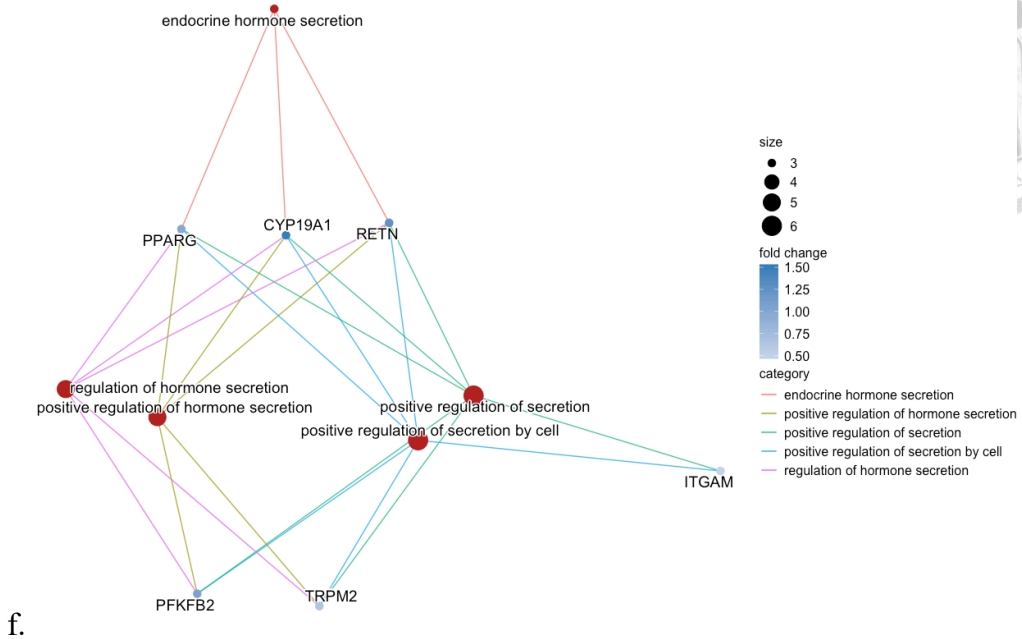


d.

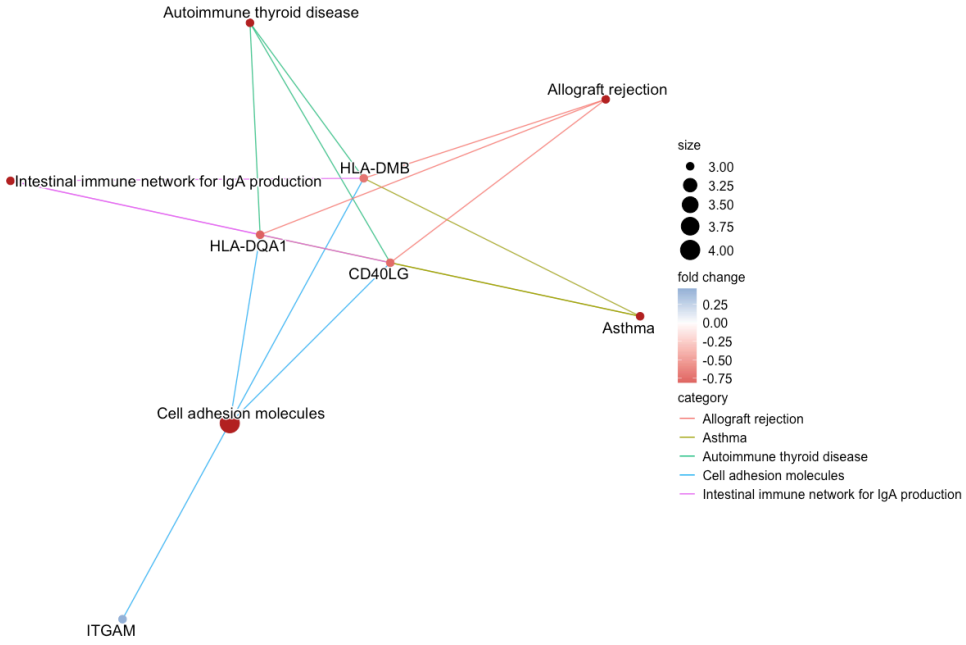


e.

doi:10.6342/NTU202402334

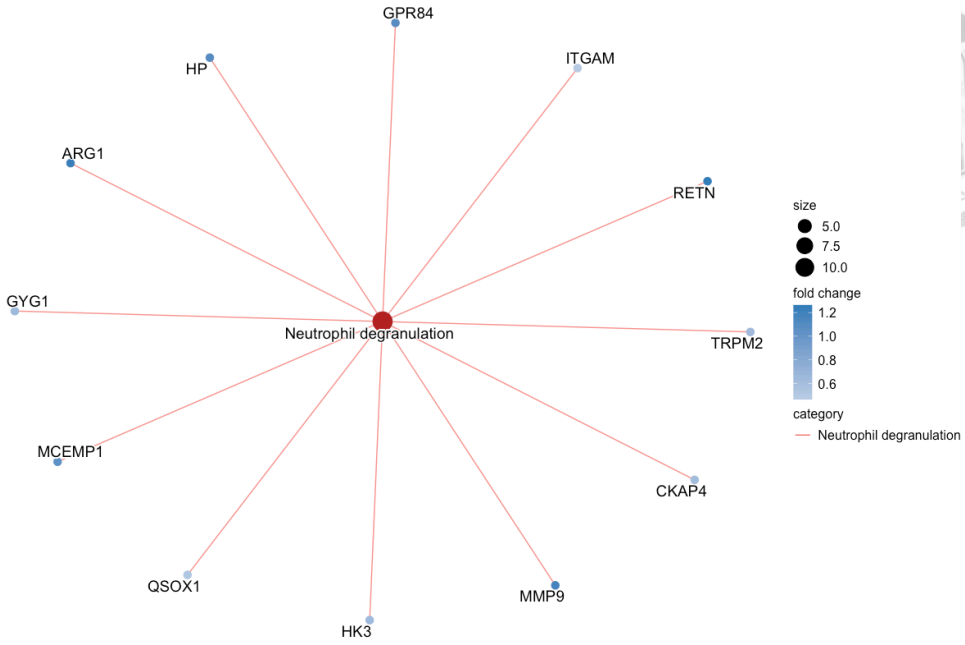


f.

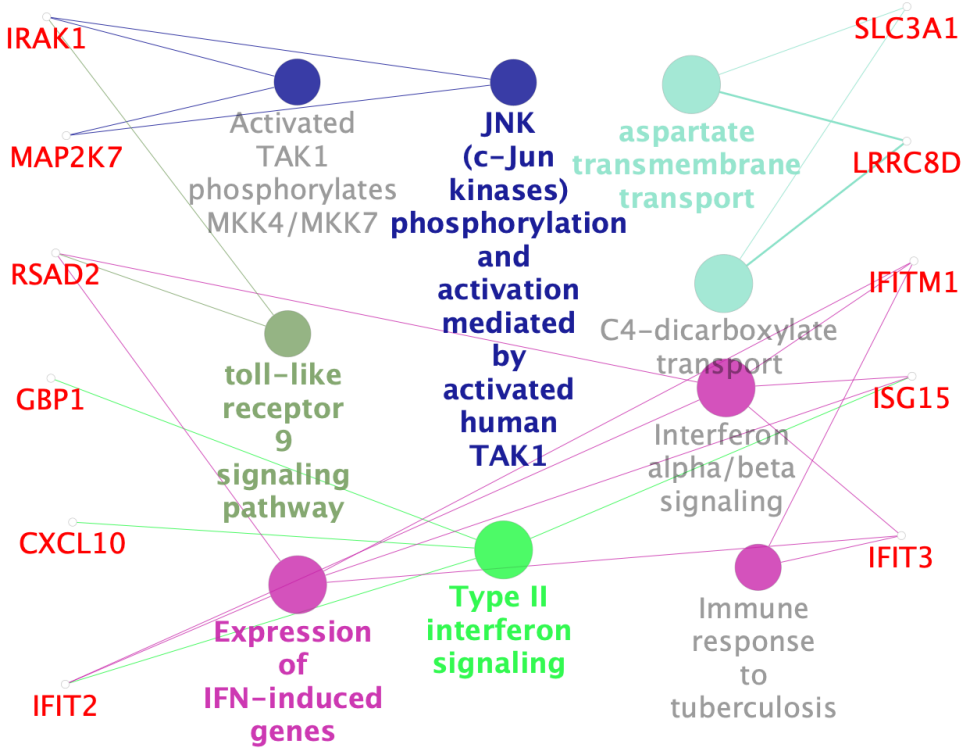


g.

doi:10.6342/NTU202402334

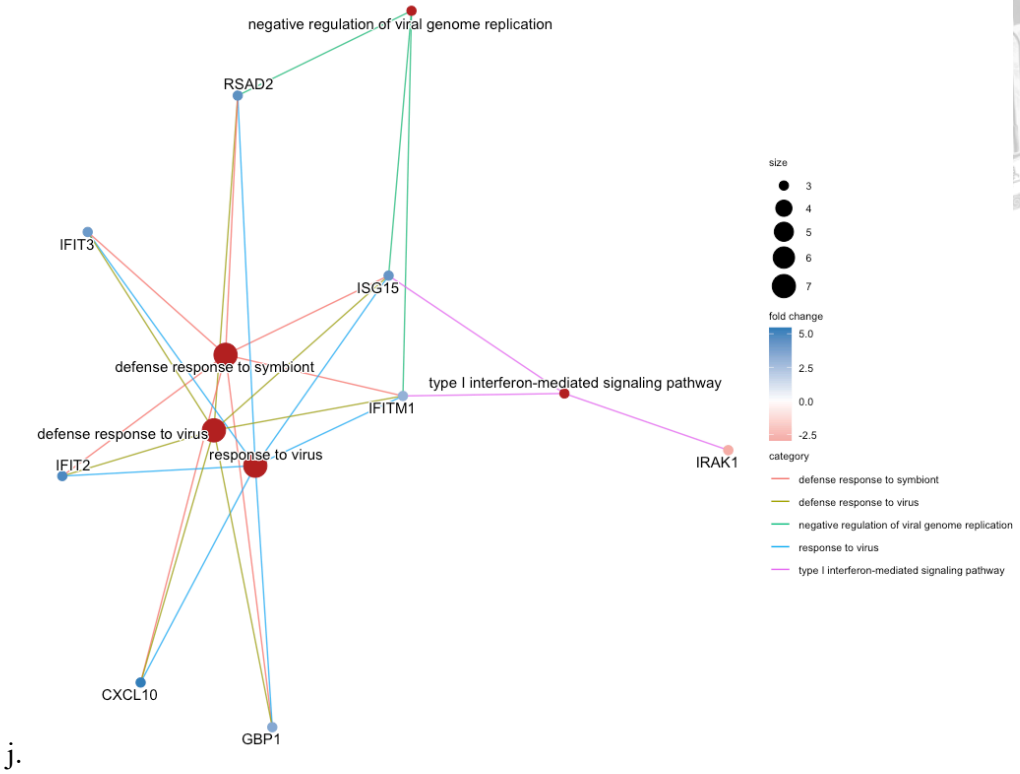


h.

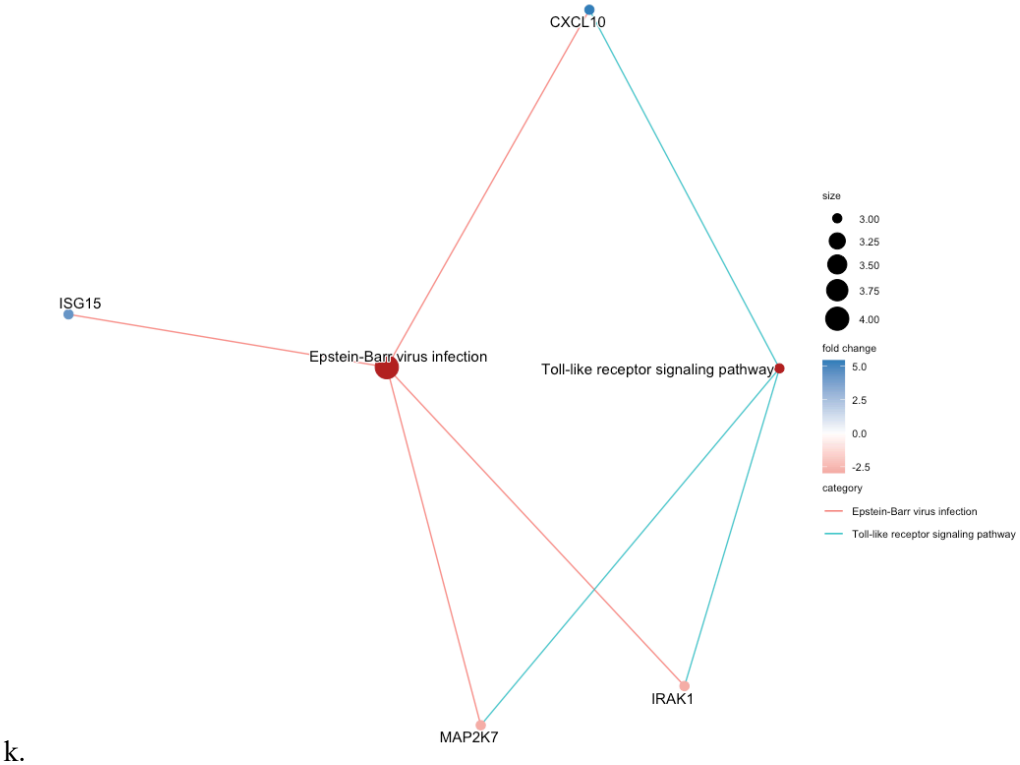


i.

doi:10.6342/NTU202402334

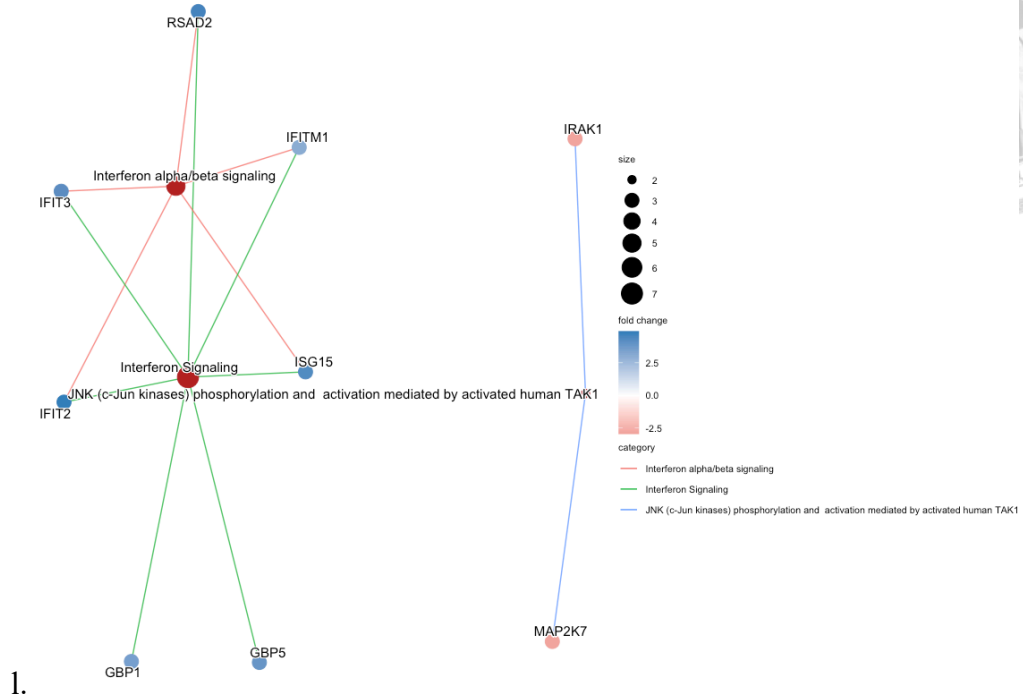


j.

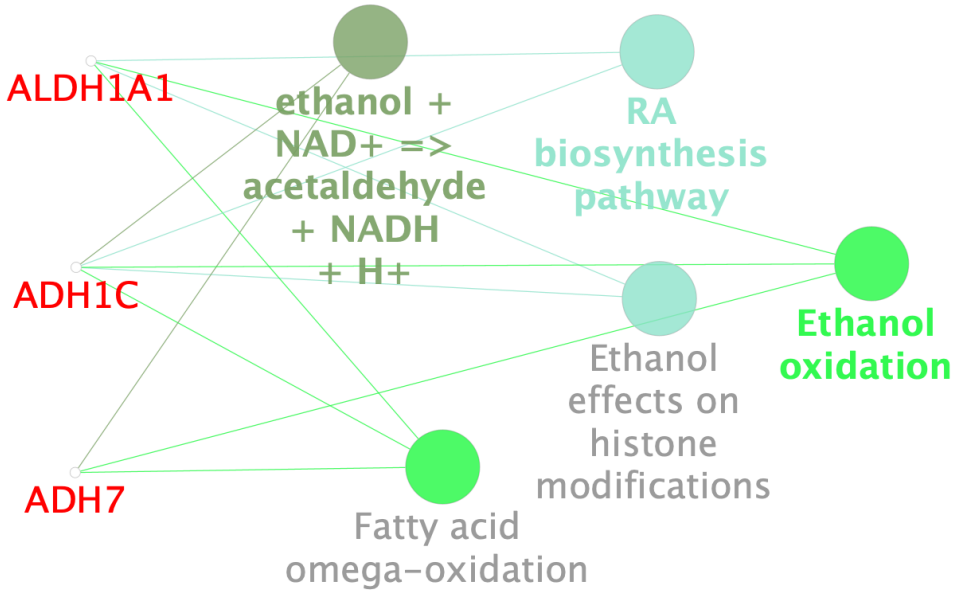


k.

doi:10.6342/NTU202402334

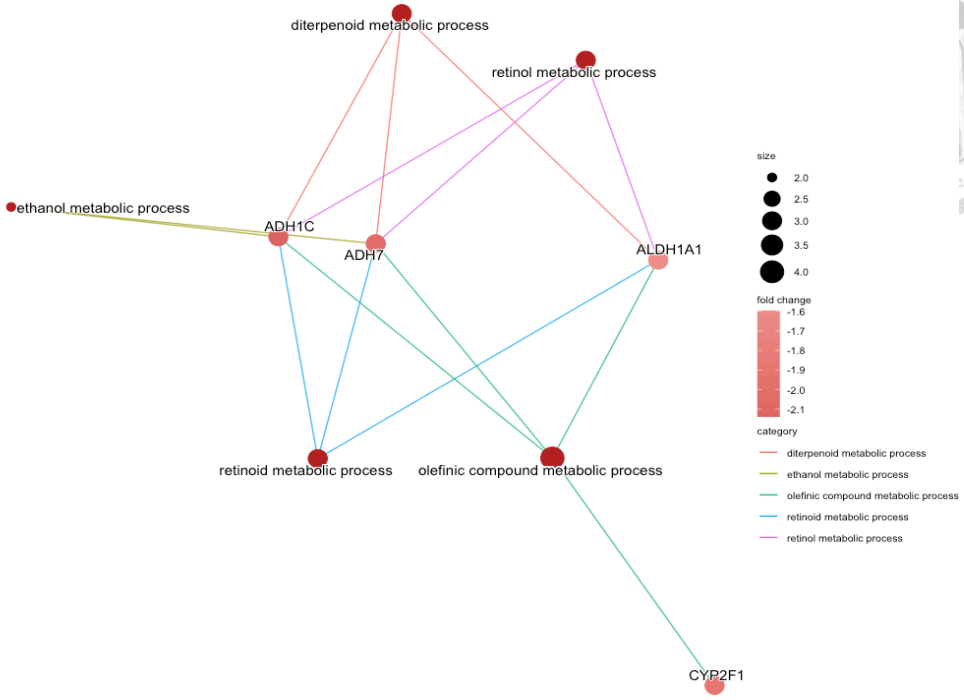


l.

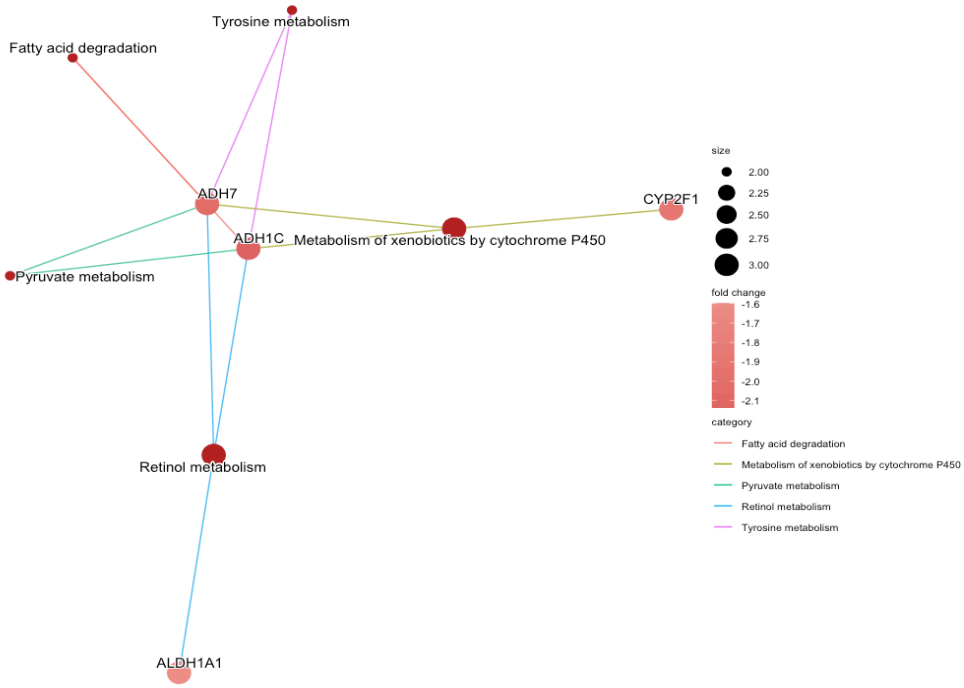


m.

doi:10.6342/NTU202402334

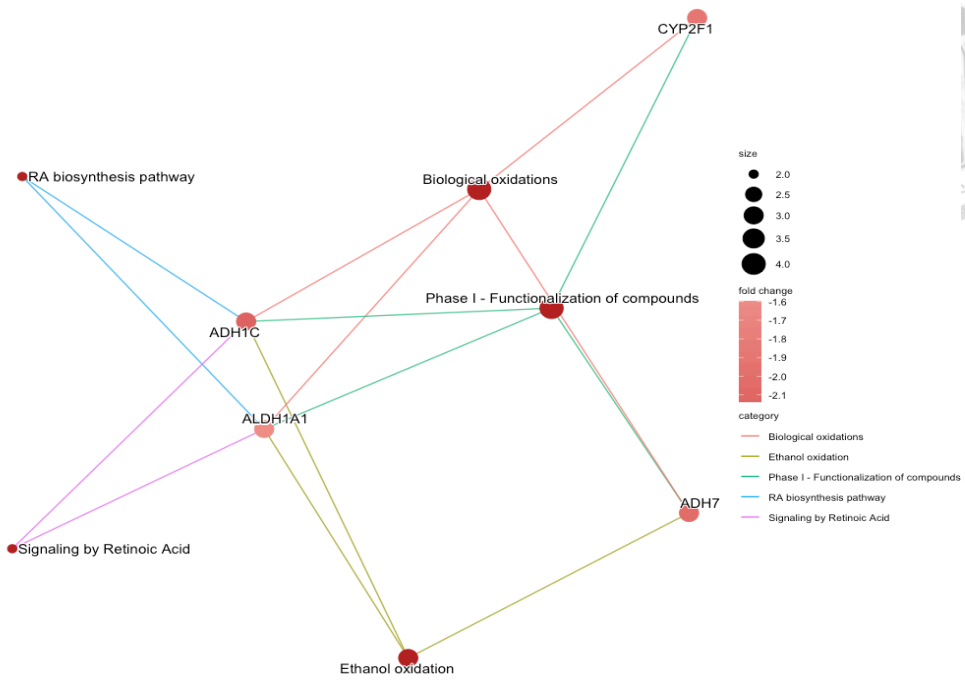


n.

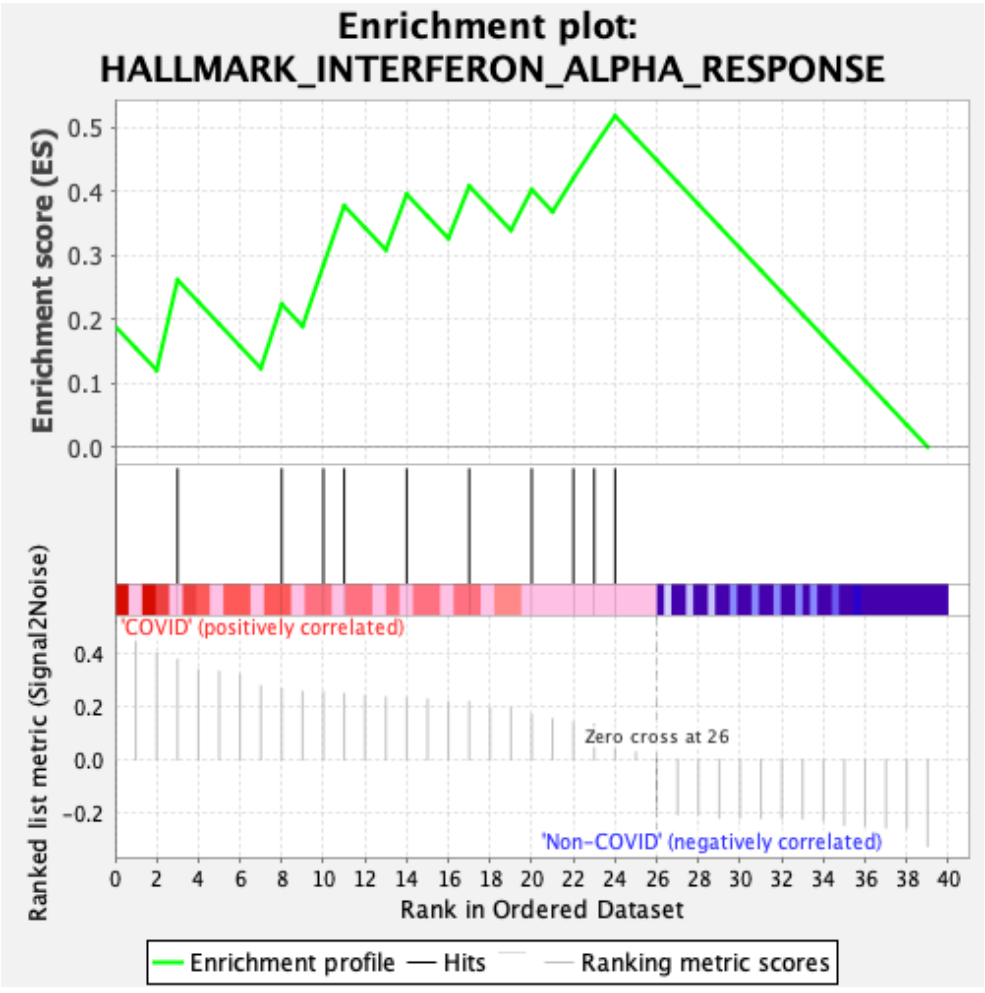


o.

doi:10.6342/NTU202402334

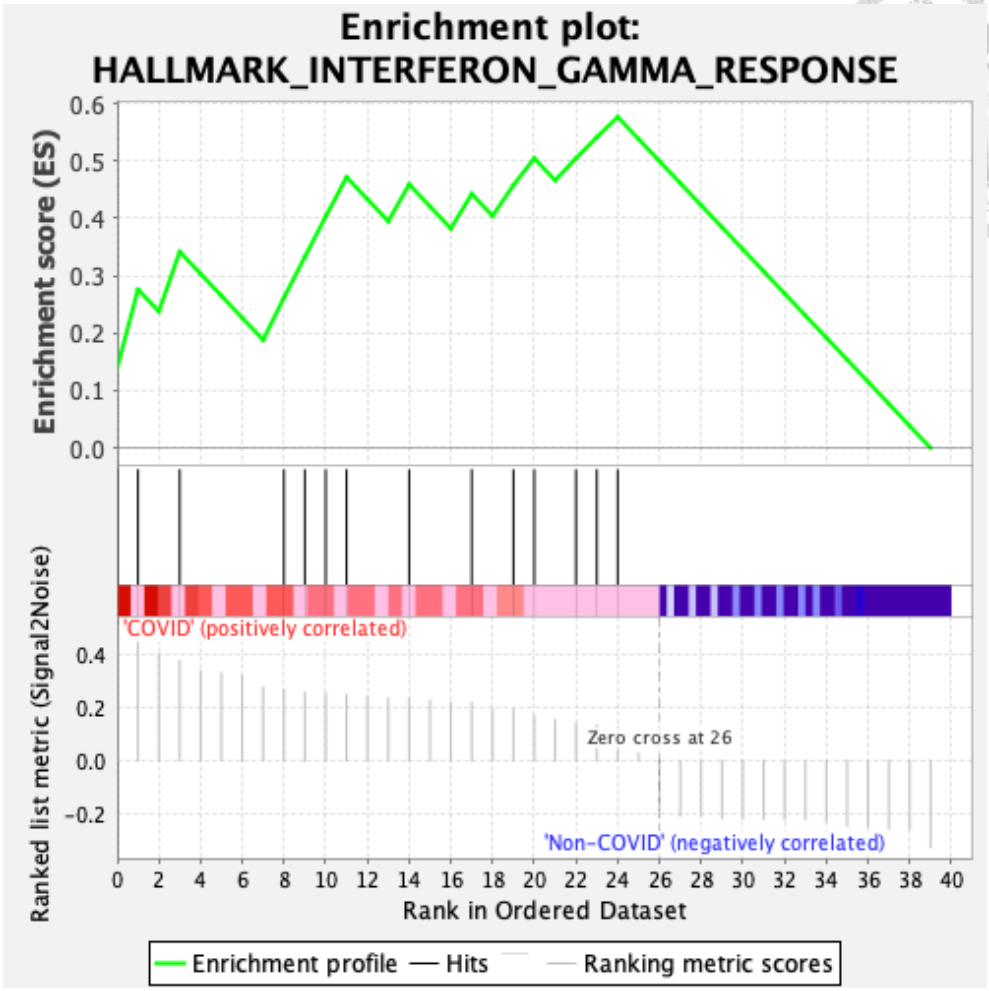


p.



q.

doi:10.6342/NTU202402334



s.

Fig 4. ClueGo integrated pathway network analysis and Cnetplot of enrichment analysis for the top 40 significantly differentially expressed genes.

(a, b, c, d) for COVID-19 vs health control in GSE152075, (e, f, g, h) GSE185263 for SOFA score >=2 vs SOFA score <2, (i, j, k, l) GSE163151 for non-COVID-19 acute respiratory infections vs health control, and (m, n, o, p) GSE163151 for non-COVID-19 acute respiratory infections vs COVID-19. (q, s) the top 40 significantly expressed genes in GSE152075 with GSEA. Reuse and modified from ref. (1).

doi:10.6342/NTU202402334

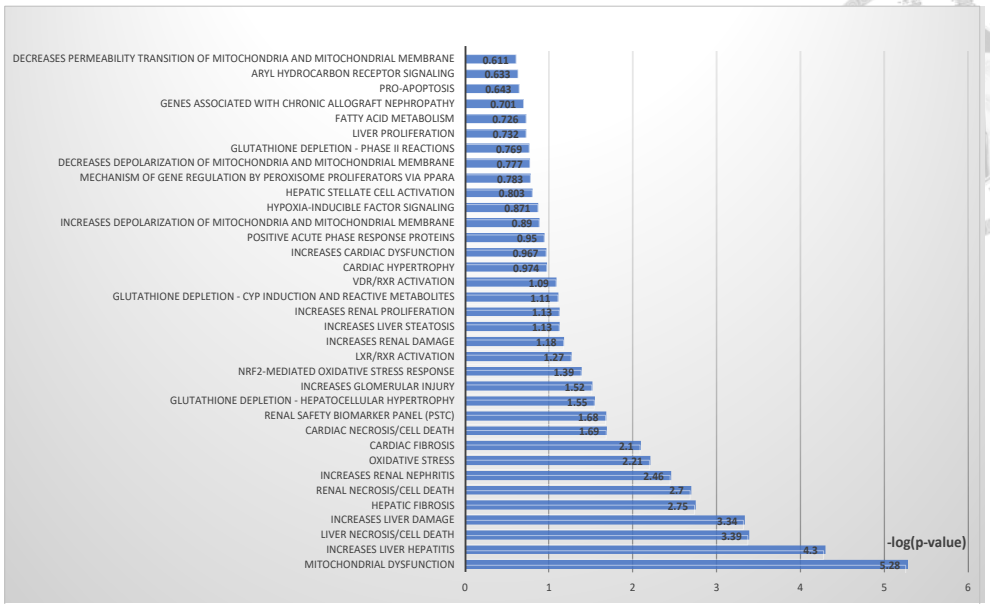
3.4 Tox and common gene analysis of heart, liver, kidney, and mitochondria for COVID-19 sample sets

The toxicity lists were generated by conducting tox analysis on the differentially expressed genes in the sample sets (Table 1) using IPA. The findings obtained from tox analysis in the nasopharyngeal swab samples of GSE152075 and GSE163151 reveal the notable toxic impacts of COVID-19 on the heart, kidney, liver, and mitochondria (Fig 5a, 5b**)**. The tox analysis of leukocyte data from GSE157103 and whole blood samples from GSE152641 revealed that COVID-19 negatively affected the heart, kidney, and liver. However, the impact on mitochondria was not as significant as GSE152075 and GSE163151. (Fig 5c, 5d**)**.

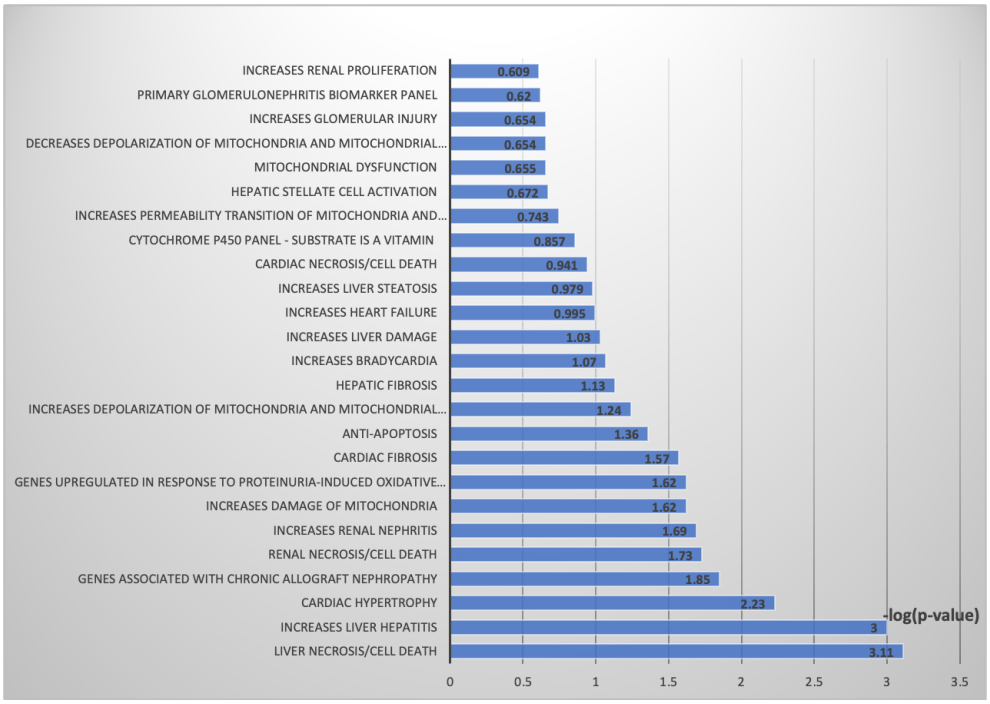
The nasopharyngeal swab data from GSE152075 exhibited the highest amount of gene overlap with nasopharyngeal swab data, which is the same tissue from GSE163151 compared to other tissue sample groups. The samples obtained from various tissues and sampling platforms exhibited distinct toxicity profiles and genes in common (Fig 5e, 5f), suggesting that the development of COVID-19 is tissue-specific.

doi:10.6342/NTU202402334

a.

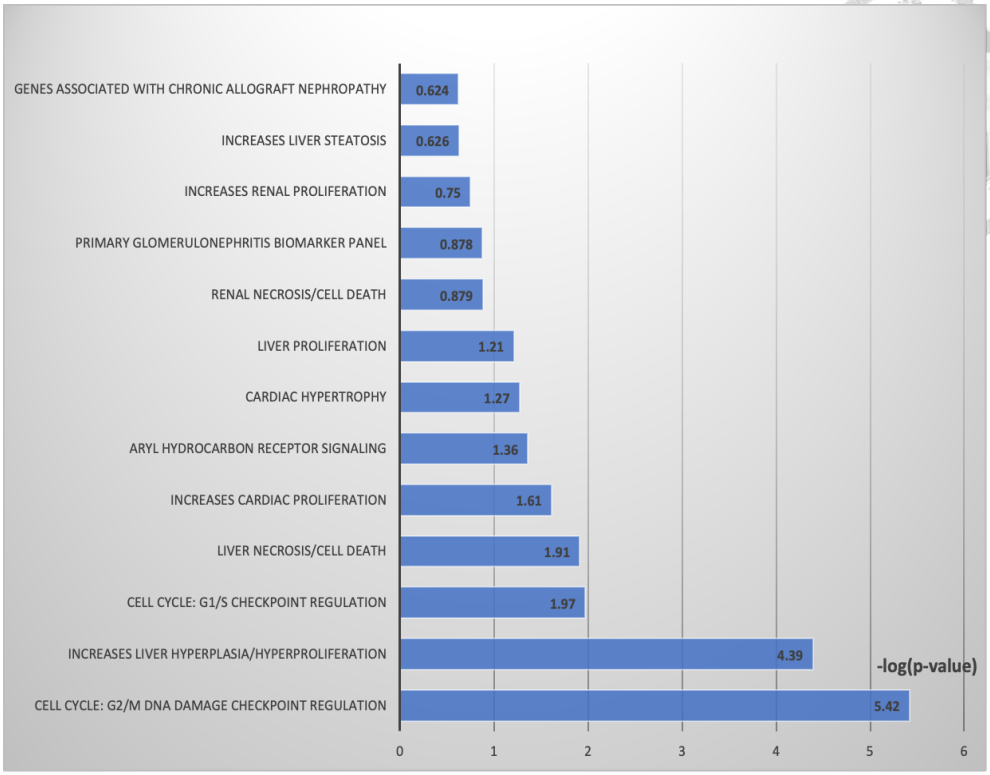


b.

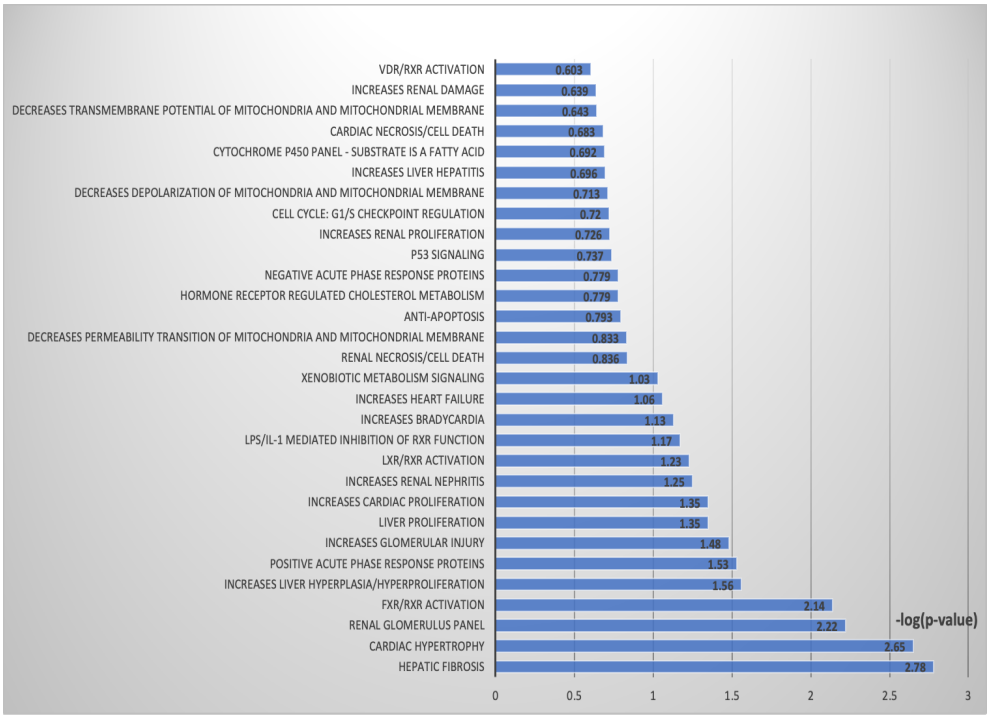


doi:10.6342/NTU202402334

c.

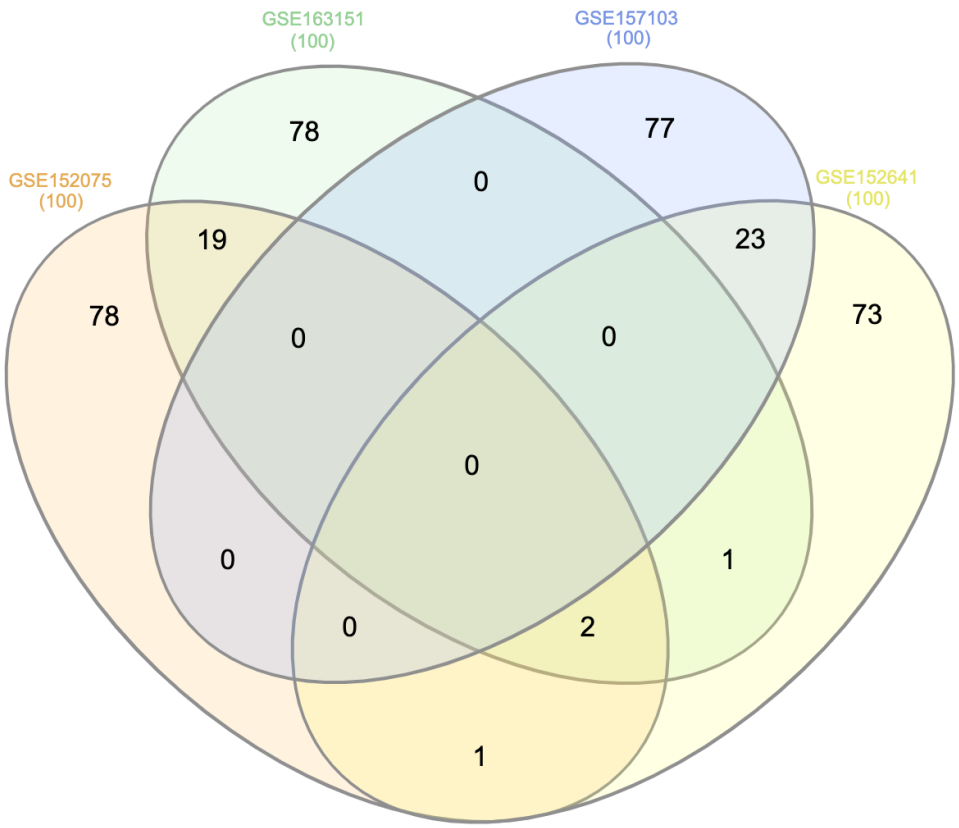


d.



doi:10.6342/NTU202402334

e.



f.

GEO accession	GSE152075 GSE152641 GSE163151	GSE152075 GSE163151	GSE152075 GSE169241	GSE152075 GSE152641
Total	2	19	1	1
Genes	*SERPING1* *EPSTI1*	*RSAD2* *MX1* *DDX60L* *TRIM22* *SAMD9L* *DDX60* *DDX58* *IFI6* *XAF1* *SAMD9* *IFI44L* *IFIT1* *GBP1* *IFITM1* *PARP9* *IFI44* *MX2* *IFIT3* *IFIT2*	*SLC8A1*	*TUBB2A*

Fig 5. Tox and common gene analysis of the RNA-seq data for COVID-19 vs health control sample sets. Reuse and modified from ref. (1).

doi:10.6342/NTU202402334

The tox analysis in IPA was conducted to generate the toxicity lists and toxicity functions of the genes that were differentially expressed in COVID-19 patients compared to health control groups. (a) toxicity lists of nasopharyngeal swab samples from GSE152075 ; (b) toxicity lists of nasopharyngeal swab samples from GSE163151; (c) toxicity lists of leukocyte samples from GSE157103, and (d) toxicity lists of whole blood samples from GSE152641; Out of the 100 significantly differentially expressed genes, the common genes that are differentially expressed across COVID-19 and the health control groups in the datasets GSE152075, GSE163151, GSE157103, and GSE152641 are identified. (e) Venn diagram. (f) common genes table. The samples collected from different tissues and sampling platforms displayed different toxicity profiles and common genes, indicating that the development of COVID-19 is tissue-specific. Reuse and modified from ref. (1).

3.5 Tox analysis for COVID-19 ICU patients, sepsis and non-COVID-19 acute respiratory infections sample sets

Nevertheless, the toxicity lists derived from the GSE157103 leukocyte samples for the analysis of toxicity in ICU vs non-ICU patients revealed the presence of toxic effects on the heart, liver, kidneys, and mitochondria (Fig 6a). Compared with no toxic mitochondria list finding in the toxicity list of GSE157103 leukocyte samples for the tox analysis in COVID-19 patients vs the health control group, this indicates that mitochondrial dysfunction is linked to the progression of COVID-19 in patients (48). And COVID-19 patients in the intensive care unit typically experience multiorgan failure.

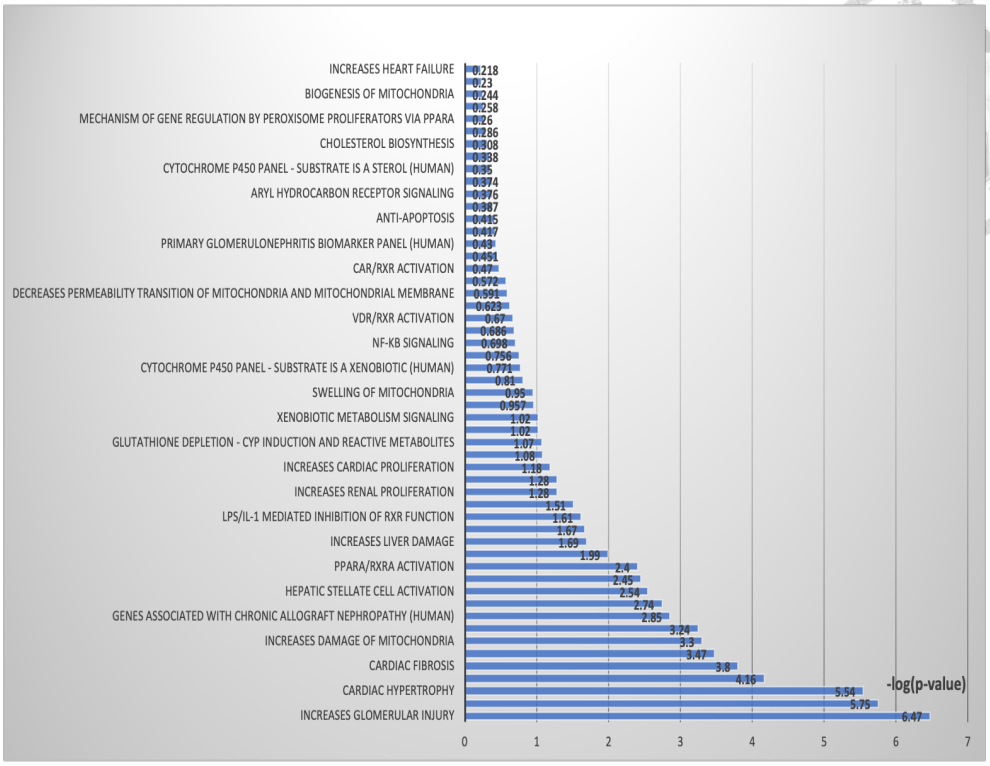
doi:10.6342/NTU202402334

To further investigate the effect of multiorgan failure, the sample data of GSE185263 for SOFA score >=2 vs. SOFA score <2 was also conducted by tox analysis. Patients with organ failure are indicated by SOFA score >=2 in sepsis. In the toxicity list of GSE185263 whole blood samples for SOFA score >=2 vs. SOFA score <2 with the tox analysis, it revealed the presence of toxic effects on the heart, kidneys, and liver, but there is no tox mitochondria exits in the toxicity list (Fig 6b). Both COVID-19 and sepsis have the potential to result in multiorgan failure. Nevertheless, the pathogenesis is different.

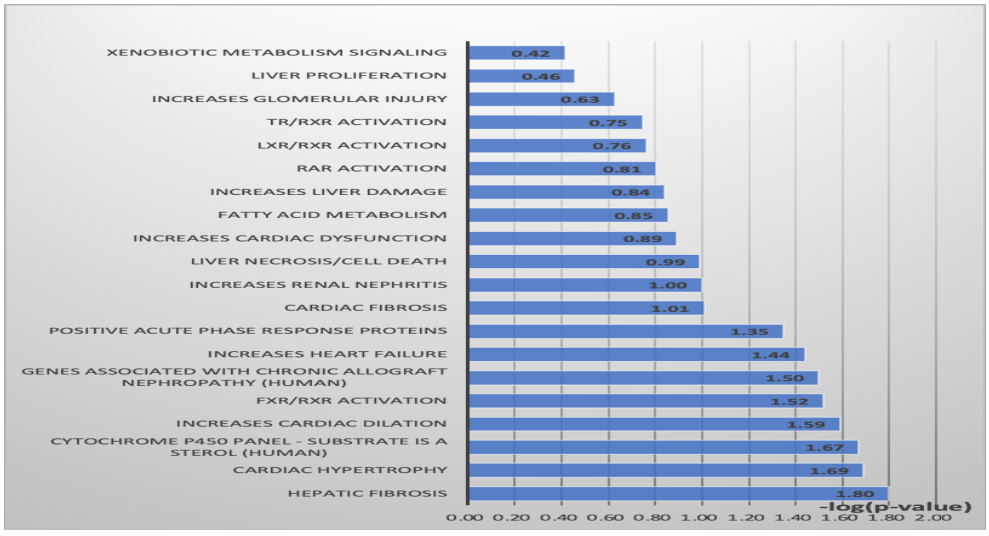
In the toxicity lists derived from the GSE163151 nasopharyngeal swab sample data for the analysis of toxicity in non-COVID-19 acute respiratory infections vs health control, the result is quite similar to COIVD-19 vs health control. It also reveals the notable toxic impacts of other non-COVID-19 acute respiratory infections on the heart, kidney, liver, and mitochondria (Fig 6c). On the other hand, the toxicity lists derived from the GSE163151 nasopharyngeal swab sample data for the analysis of toxicity in non-COVID-19 acute respiratory infections vs COVID-19, mitochondria is still play an important role in the toxicity lists (Fig 6d). This topic requires additional exploration in relation to long COVID, as there is a higher prevalence of patients experiencing long COVID symptoms in COVID-19 compared to other non-COVID-19 acute respiratory infections.

doi:10.6342/NTU202402334

a.

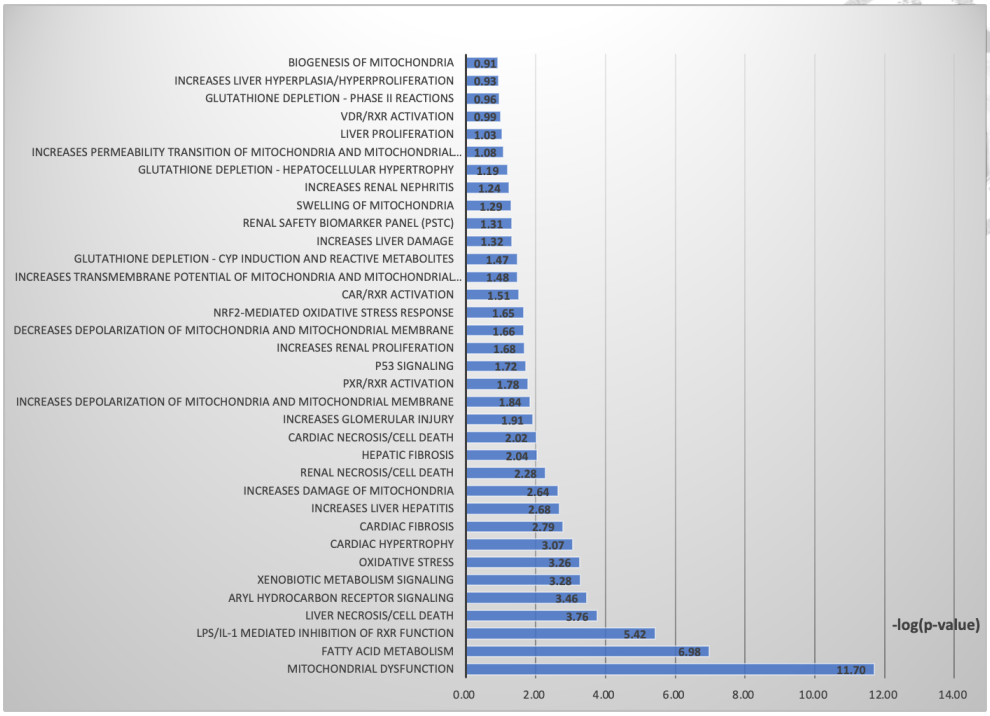


b.



doi:10.6342/NTU202402334

c.



d.

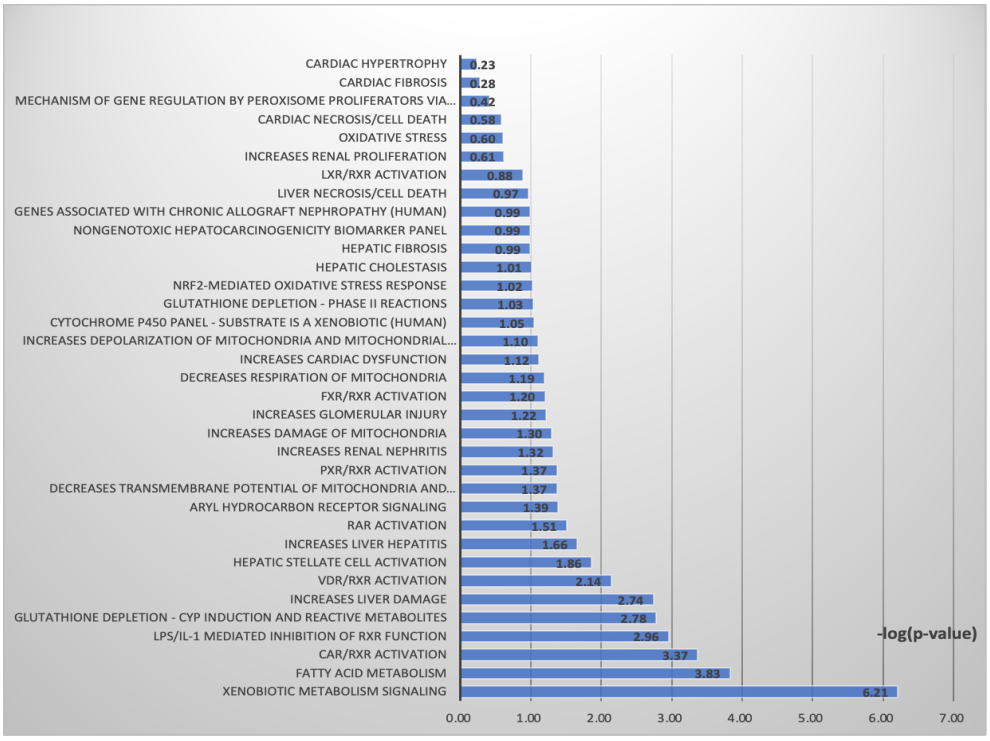


Fig 6. Tox analysis of the RNA-seq data for COVID-19, sepsis and other non-COVID-19 acute respiratory infections sample sets.

The tox analysis in IPA was conducted to generate the toxicity lists and toxicity functions of the genes that were differentially expressed in COVID-19 ICU vs non-ICU,

doi:10.6342/NTU202402334

sepsis SOFA score >=2 vs SOFA score <2, and non-COVID-19 acute respiratory infections vs health control/COVID-19. (a) ICU patients and non-ICU patients were compared in GSE157103 leukocyte sample set. (b) GSE185263 whole blood sample data for SOFA score >=2 vs SOFA score <2. (c) GSE163151 nasopharyngeal swab sample data for non-COVID-19 acute respiratory infections vs health control. (d) GSE163151 nasopharyngeal swab sample data for non-COVID-19 acute respiratory infections vs COVID-19. Reuse and modified from ref. (1).

3.6 Machine learning for genes associated with heart-, liver-, kidney-, and mitochondria-related toxicity lists in COVID-19 and sepsis sample data

The differentially expressed genes (DEGs) of the GSE152075 sample set associated with organs-related toxicity lists from IPA-tox analysis under the threshold of -log(*p* value) > 1.3 were identified for further analysis of machine learning (Fig 5a**)**. The significantly differentially expressed genes were identified from the heart-, liver-, kidney-, and mitochondria-related toxicity lists. The heart-related toxicity lists included cardiac necrosis/cell death pathways (49), and cardiac fibrosis (50); there were 38 genes in these toxicity lists. The liver-related toxicity lists included liver necrosis/cell death (51), increases in liver hepatitis (52), and increases in liver damage pathways (53); there were 42 genes in these toxicity lists. The kidney-related toxicity list included increases glomerular injury pathways (54), increases renal nephritis (55) panel (PSTC) (56), and renal necrosis/cell death (57); there were 55 genes in these toxicity lists. Similarly, mitochondrial dysfunction was the relevant pathway identified by the mitochondria-related toxicity list, which included 32 DEGs in GSE150275 (Table 7). The genes

doi:10.6342/NTU202402334

identified from the toxicity lists were compared to the 40 genes that showed substantial differential expression. Only a small number of genes were found to be both significantly differentially expressed and common between the two sets, including *RRAD* in heart-, *CXCL10* in liver-, and *PRDX5* in mitochondria-related toxicity list. Significant disparities in gene expression were observed between the groups defined by the statistical significance method and those identified through the biological significance method.

Table 7. Significantly differentially expressed genes associated with heart-, liver-, kidney-, and mitochondria-related toxicity list and in GSE152075. Reuse and modified from ref. (1).

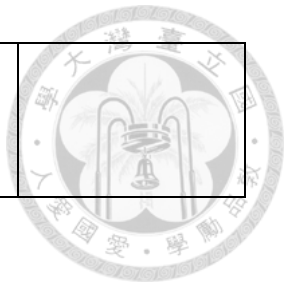
	Genes associated with heart-related toxicity list (38 genes)	Genes associated with liver-related toxicity list (42 genes)	Genes associated with renal-related toxicity list (55 genes)	Genes associated with mitochondrial-related toxicity list (32 genes)
Organs-related toxicity list	Cardiac Fibrosis Cardiac Necrosis/Cell Death	Increases Liver Hepatitis Liver Necrosis/Cell Death Increases Liver Damage	Renal Necrosis/Cell Death Increases Renal Nephritis Renal Safety Biomarker Panel (PSTC) Increases Glomerular Injury	Mitochondrial Dysfunction

doi:10.6342/NTU202402334

Genes in organs-related toxicity list	*CYBB, TLR4, PRKCB, SLC8A1, TLR2, GPX1, ACE2, DVL1, SLC2A1, STEAP3, CIB1, NDUFS6, MB, CARD6, JUN, LMNA, KLF15, LARP6, DAG1, USP18, NEXN, FLT1, PIN1, NUB1, CTNNB1, PROX1, S100A6, CCN1, RRAD, CASP1, BAD, NDUFA13, SOCS3, KLF4, NTN1, SERPINF1, JUND, ABCC9*	*TLR4, GBP5, CASP1, CCL4, TLR2, TLR7, JUN, TNFSF14, ALDH3A1, P2RX7, KMO, MIF, BSG, CCR5, IL2RG, ATG4B, AIM2, CCN1, EPHA2, CCL2, KEAP1, CHCHD2, PROS1, SELL, KRT8, CTNNB1, FOS, CXCL10, IRF8, CD274, TKT, BAD, FGL2, GADD45B, SIGIRR, SOCS3, IER3, BIRC3, USP18, PTPRC, JUND, PHB2*	*MIF, TLR2, CLU, CTNNB1, FLT1, TLR7, TNFSF13B, TNFSF14, OLR1, LRP5, CYBB, GPX4, ZEB1, TLR4, P2RX7, KMO, ACE2, SLC2A1, FOS, LAMB2, PRKCB, MEFV, APRT, HSPA1A, HSPA1B, NDUFAB1, CASP1, STUB1, BAD, ALDH3B1, MLKL, BCL2L14, RFXANK, IDO1, SLC8A1, PRDX2, IER3, GNB2, BIRC3, CITED2, NTN1, GSTP1, ERBB2, ZBP1, APOBEC3A, AIM2, FCGR3A, FCGR3B, CX3CR1, TFF3, DDR1, CD274, JUN, C3AR1, CCR1*	*ACO2, ATP5F1E, ATP5ME, BAD, CLIC2, COX4I1, COX5A, COX5B, COX6A1, COX7B, CYC1, FIS1, GPX1, GPX4, GSTP1, MT-ND6, NDUFA11, NDUFA13, NDUFA2, NDUFA4, NDUFAB1, NDUFB10, NDUFB2, NDUFB7, NDUFS6, NDUFV1, PRDX5, SURF1, TOMM7, UQCR11, UQCRC1, UQCRQ*
Common genes with top 40 significantly differentially expressed	*RRAD*	*CXCL10*	_	*PRDX5*

doi:10.6342/NTU202402334

genes from GSE152075				



To use the top 40 significantly differentially expressed genes of sepsis status to predict COVID-19 and non-COVID-19 status, the top 40 significantly differentially expressed genes of GSE185263 was be utilized in the sample GSE152641 which is the tissue of whole blood and GSE152075 which is the tissue of nasopharyngeal swab for machine learning analysis. However, in the sample data GSE152641, the machine learning performance of the top 40 significantly differentially expressed genes from GSE185263 to predict COVID-19 and non-COVID-19 is superior compared to GSE152075. It is because the tissue sampled in GSE185263 and GSE152641 is whole blood, whereas GSE152075 is nasopharyngeal swab. Although the tissue sampled in GSE185263 is whole blood, whereas GSE152075 involves a nasopharyngeal swab, this machine learning investigation still provides evidence that the severity of sepsis can be used as a prediction indication for the severity of COVID-19 (46). Hence, the significantly differentially expressed genes from sepsis could be the classifier of COVID-19.

On the other hand, the genes that were expressed differently in relation to heart, liver, kidney, and mitochondria toxicity were utilized as features to evaluate the predictive capabilities of various machine learning models for COVID-19. The findings indicated that the gene sets associated with heart, liver, kidney, and mitochondria toxicity were able to achieve high levels of accuracy, F1-Score, and AUC for the four machine learning methods, except for SVM, which had a performance level of 90% or higher (Table 8). The machine learning analysis results illustrated that the differentially expressed genes in the toxicity list of the heart, liver, kidney, and mitochondria were

doi:10.6342/NTU202402334

correlated with COVID-19 and were sufficient for machine learning to predict the disease.

Table 8. Machine learning results of top 40 significantly differentially expressed genes from GSE185263 in sample data GSE185263 and GSE152075; The genes associated with the heart-, liver-, kidney-, and mitochondria-related toxicity list in samples data GSE152075. Reuse and modified from ref. (1).

Machine Learning \ Feature		40 genes from GSE185263 with GSE152641	40 genes from GSE185263 with GSE152075	Genes associated with heart-related toxicity list (38 genes)	Genes associated with liver-related toxicity list (42 genes)	Genes associated with renal-related toxicity list (55 genes)	Genes associated with mitochondrial-related toxicity list (32 genes)
XGBoost	Accuracy	0.884	0.919	0.938	0.942	0.938	0.948
	Sensitivity	0.919	0.977	0.970	0.967	0.974	0.981
	Specificity	0.792	0.463	0.685	0.741	0.648	0.685
	Precision	0.919	0.935	0.961	0.967	0.957	0.961
	F1-Score	0.919	0.956	0.965	0.967	0.965	0.971
	MCC	0.711	0.534	0.678	0.708	0.668	0.723
	AUC	0.948	0.814	0.913	0.933	0.913	0.859
Random Forest	Accuracy	0.895	0.928	0.952	0.944	0.940	0.944
	Sensitivity	0.919	0.986	0.993	0.981	0.986	0.991
	Specificity	0.833	0.463	0.630	0.648	0.574	0.574
	Precision	0.934	0.936	0.955	0.957	0.949	0.949
	F1-Score	0.927	0.960	0.974	0.969	0.967	0.969
	MCC	0.744	0.577	0.738	0.697	0.664	0.687
	AUC	0.958	0.865	0.955	0.969	0.960	0.922

doi:10.6342/NTU202402334

Logistic Regression	Accuracy	0.930	0.862	0.897	0.909	0.907	0.905
	Sensitivity	0.952	0.888	0.907	0.914	0.912	0.930
	Specificity	0.875	0.648	0.815	0.870	0.870	0.704
	Precision	0.952	0.953	0.975	0.982	0.982	0.962
	F1-Score	0.952	0.919	0.940	0.947	0.946	0.946
	MCC	0.827	0.448	0.600	0.652	0.647	0.574
	AUC	0.963	0.785	0.898	0.939	0.919	0.863
SVM	Accuracy	0.930	0.864	0.899	0.723	0.771	0.936
	Sensitivity	0.935	0.909	0.919	0.693	0.756	0.979
	Specificity	0.917	0.500	0.741	0.963	0.889	0.593
	Precision	0.967	0.935	0.966	0.993	0.982	0.95
	F1-Score	0.951	0.922	0.942	0.816	0.854	0.964
	MCC	0.832	0.376	0.574	0.425	0.437	0.646
	AUC	0.940	0.868	0.949	0.966	0.937	0.935

Adding the significantly differentially expressed genes of mitochondria-related toxicity lists to the top 40 significantly differentially expressed genes related to multiorgan damage caused by sepsis for machine learning. This aims to investigate if mitochondrial dysfunction will worsen multiorgan damage and further improve the prediction performance of machine learning as a classifier for COVID-19.

Additionally, to determine whether COVID-19 may exacerbate cardiac, hepatic, and renal function failure as a result of mitochondrial dysfunction, the significantly differentially expressed genes of mitochondria-related toxicity lists were incorporated into the significantly differentially expressed genes related to heart, liver, and kidney toxicity list.

Upon merging the genes in the lists, a total of 72 genes exhibited significant differential expression in relation to both the mitochondria and the multiorgan damage

doi:10.6342/NTU202402334

caused by sepsis, 66 genes exhibited significant differential expression in relation to both the mitochondria and the heart, 83 genes exhibited significant differential expression in relation to both the mitochondria and the kidney, and 73 genes exhibited significant differential expression in relation to both the mitochondria and the liver.

The results showed that by adding genes in the mitochondria-related toxicity list to multiorgan damage caused by the sepsis genes set, the machine learning performance would improve, especially in logistic regression and SVM.

Furthermore, including the genes expressed significantly differentially in the mitochondria-related toxicity list, together with the heart-, liver-, and kidney-related toxicity lists, enhanced predictive capabilities in machine learning models (Table 9).

Table 9. Machine learning results of significantly differentially expressed genes associated with mitochondria dysfunction plus multiorgan damage caused by sepsis, heart-, liver-, and kidney-related toxicity lists. Reuse and modified from ref. (1).

Machine Learning \ Features		Genes from mitochondrial dysfunction + 40 genes from GSE185263 with GSE152641 (72 genes)	Genes from mitochondrial dysfunction + heart-related toxicity list (66 genes)	Genes from mitochondrial dysfunction + liver-related toxicity list (73 genes)	Genes from mitochondrial dysfunction + renal-related toxicity list (83 genes)
XGBoost	Accuracy	0.895	0.955	0.934	0.946
	Sensitivity	0.903	0.986	0.965	0.984
	Specificity	0.875	0.704	0.685	0.648

doi:10.6342/NTU202402334

	Precision	0.949	0.964	0.961	0.957
	F1-Score	0.926	0.975	0.963	0.970
	MCC	0.752	0.755	0.661	0.707
	AUC	0.970	0.905	0.952	0.912
Random Forest	Accuracy	0.895	0.957	0.963	0.957
	Sensitivity	0.919	0.991	0.998	0.993
	Specificity	0.833	0.685	0.685	0.667
	Precision	0.934	0.962	0.962	0.960
	F1-Score	0.927	0.976	0.979	0.976
	MCC	0.744	0.764	0.799	0.763
	AUC	0.971	0.957	0.970	0.964
Logistic Regression	Accuracy	0.965	0.919	0.940	0.940
	Sensitivity	0.968	0.933	0.951	0.953
	Specificity	0.958	0.815	0.852	0.833
	Precision	0.984	0.976	0.981	0.979
	F1-Score	0.976	0.954	0.966	0.966
	MCC	0.915	0.657	0.732	0.727
	AUC	0.968	0.906	0.953	0.922
SVM	Accuracy	0.930	0.917	0.913	0.911
	Sensitivity	0.984	0.935	0.926	0.926
	Specificity	0.792	0.778	0.815	0.796
	Precision	0.924	0.971	0.975	0.973
	F1-Score	0.953	0.953	0.950	0.949
	MCC	0.823	0.638	0.641	0.628
	AUC	0.949	0.952	0.969	0.936

Including mitochondria dysfunction-related genes in the heart, liver, and kidney-related toxicity lists resulted in a substantial improvement in the performance indices of machine learning. More specifically, logistic regression and SVM demonstrated a

doi:10.6342/NTU202402334

considerable improvement when comparing machine learning using the top 40 significant genes in GSE152075 and the significant genes related to mitochondrial dysfunction only (Fig 7). Hence, this outcome implies that COVID-19 might exacerbate cardiac, liver, and kidney dysfunction by inducing mitochondrial damage.

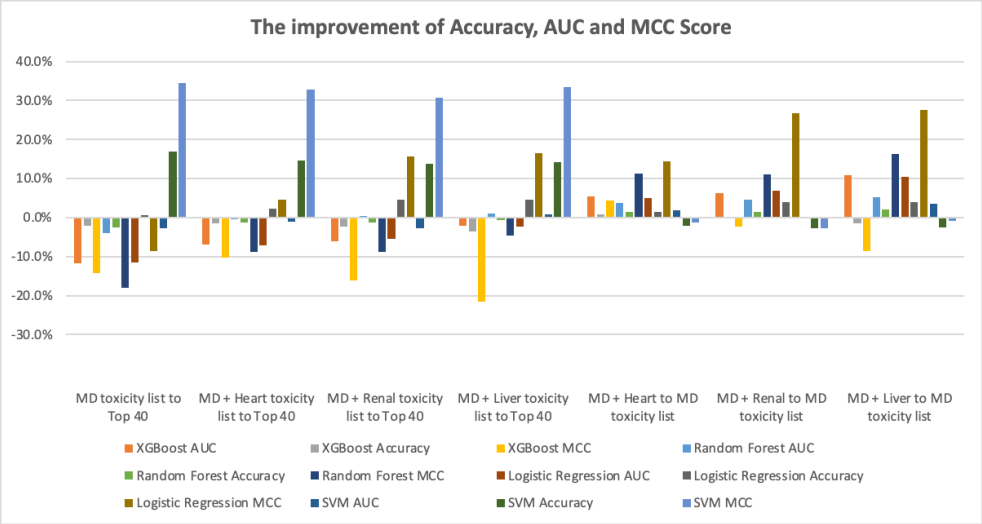


Fig 7. Machine learning performance comparison by adding mitochondrial-related genes.

"MD" represents mitochondrial dysfunction, and the "top 40" represent the top 40 significant genes of GSE152075. Reuse and modified from ref. (1).

3.7 Analysis of the common genes associated with heart, liver, kidney, and mitochondria toxicity

Given that COVID-19 might exacerbate cardiac, liver, and renal failure by harming the mitochondria, it is important to examine the common genes among the significantly differentially expressed genes found in the mitochondria-related toxicity list and the

doi:10.6342/NTU202402334

significantly differentially expressed genes of top 40 genes from sepsis, genes from heart-, liver-, and kidney-related toxicity lists.

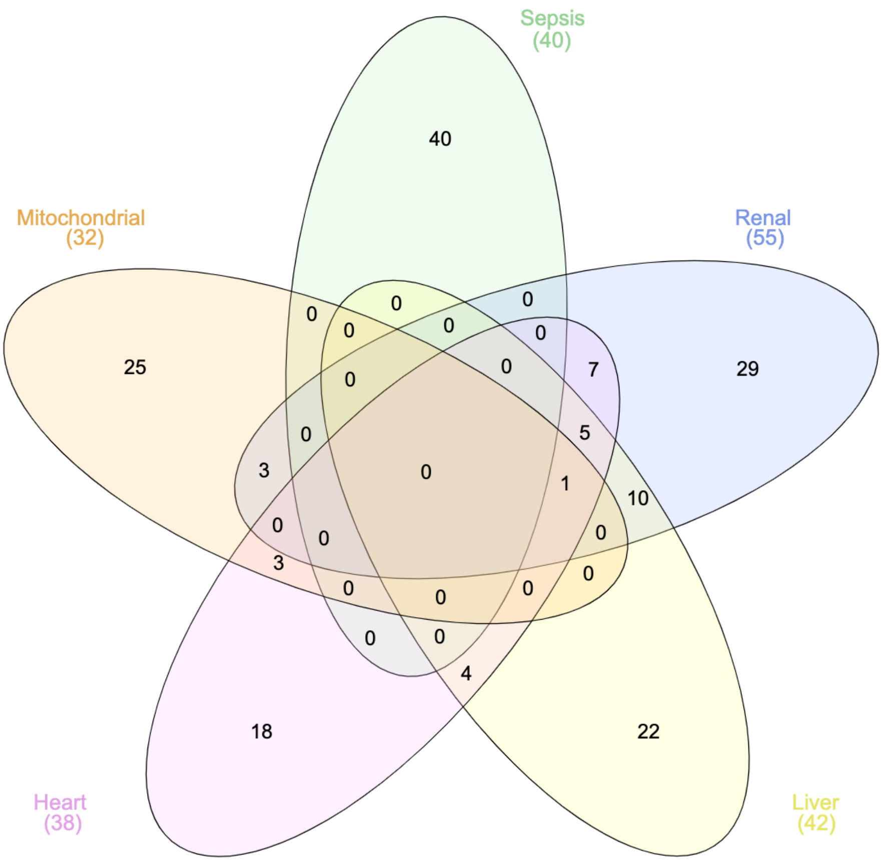
There is no common gene between the top 40 genes from sepsis and genes from mitochondria-related toxicity lists. This could be interpreted that mitochondrial dysfunction is not the main factor for the multiorgan damage caused by sepsis. The common genes between the mitochondria- and kidney-related toxicity lists were *GSTP1*, *GPX4*, *NDUFAB1*, and *BAD*; the common genes between the mitochondria- and heart-related toxicity lists were *NDUFA13, BAD, GPX1,* and *NDUFS6*; and the common genes between the mitochondria- and liver-related toxicity lists was *BAD only* (Fig 8a, 8b).

NDUFAB1 (NADH:ubiquinone oxidoreductase subunit AB1), *NDUFS6* (NADH:ubiquinone oxidoreductase subunit S6), and *NDUFA13* (NADH:ubiquinone oxidoreductase subunit A13) are genes encoding subunits of the mitochondrial complex I (NADH: ubiquinone oxidoreductase). This complex is a crucial component of the mitochondrial electron transport chain, which is responsible for ATP production through OXPHOS (oxidative phosphorylation). *GPX1* (glutathione peroxidase 1) and *GPX4* (glutathione peroxidase 4) have been identified to be associated with oxidative stress. *GPX1* and *GPX4* encode antioxidant enzymes that protect cells from oxidative stress and damage. *GSTP1* (glutathione S-Transferase Pi 1) is intricately associated with the regulation of cellular oxidative stress, inhibition of cellular apoptosis, and enhancement of cytotoxic metabolism (58). *BAD* (BCL2-associated agonist of cell death) is the sole overlapping gene among the lists of genes associated with toxicity in the heart, liver, and kidney, as well as the list of genes connected to toxicity in the mitochondria. The BAD protein, a member of the Bcl-2 gene family, is a proapoptotic factor that plays a role in starting programmed cell death (apoptosis). This may account

doi:10.6342/NTU202402334

for the cell death observed in different tissue types and its contribution to the development of diseases and organ damage (59).

Alongside mitochondrial dysfunction, the toxicity lists of GSE152075 also encompass oxidative stress and the NRF2-mediated oxidative stress response. Both oxidative stress and the NRF2-mediated oxidative stress response are highly associated with OXPHOS which produces reactive oxygen species (ROS) (60). Excessive ROS in the regulation of intracellular signaling can result in permanent harm to cellular components and trigger apoptosis by enhancing the intrinsic apoptotic pathway of the mitochondria (61). Thus, oxidative stress can induce apoptosis through a mitochondria-dependent pathway (62) which further results in multiorgan damage.



a.

doi:10.6342/NTU202402334

b.

Toxicity lists	mitochondria and top 40 genes from sepsis	mitochondria and liver	mitochondria and heart	mitochondria and renal
Total	0	1	4	4
Genes	–	*BAD*	*GPX1, NDUFA13, NDUFS6, BAD*	*GPX4, GSTP1, NDUFAB1, BAD*

Fig 8. Common genes between sepsis, other organs- and mitochondria-related toxicity lists. Reuse and modified from ref. (1).

3.8 Feature importance analysis

To assess the importance of differentially expressed genes in the lists associated with heart, liver, kidney, and mitochondria toxicity, SHAP was based on the machine learning of XGBoost to further identify the top-ranking genes. The top 3 feature importance ranking of differentially expressed genes was revealed, namely *CXCL10*, *ATP5F1E*, and *ACE2* (Fig 9a). *CXCL10* (C-X-C motif chemokine ligand 10) which encodes a cytokine known as IP-10 (interferon gamma-induced protein 10) and belongs to the CXC chemokine family. It plays a crucial role in the regulation of the immune system and inflammatory reactions. IP-10 has been shown to be an early indicator of the likelihood of severe organ failure and death in COVID-19 pneumonia (63). *CXCL10* is a significant chemokine associated with the development of cytokine storm in individuals infected with COVID-19 (64, 65). Moreover, it has been demonstrated to be a remarkable prognostic indicator for patients with COVID-19 (66, 67). The SHAP feature importance analysis exactly exhibits and validate the impact of *CXCL10* on COVID-19. On the other hand, *CXCL10* could induce mitochondrial depolarization in pancreatic

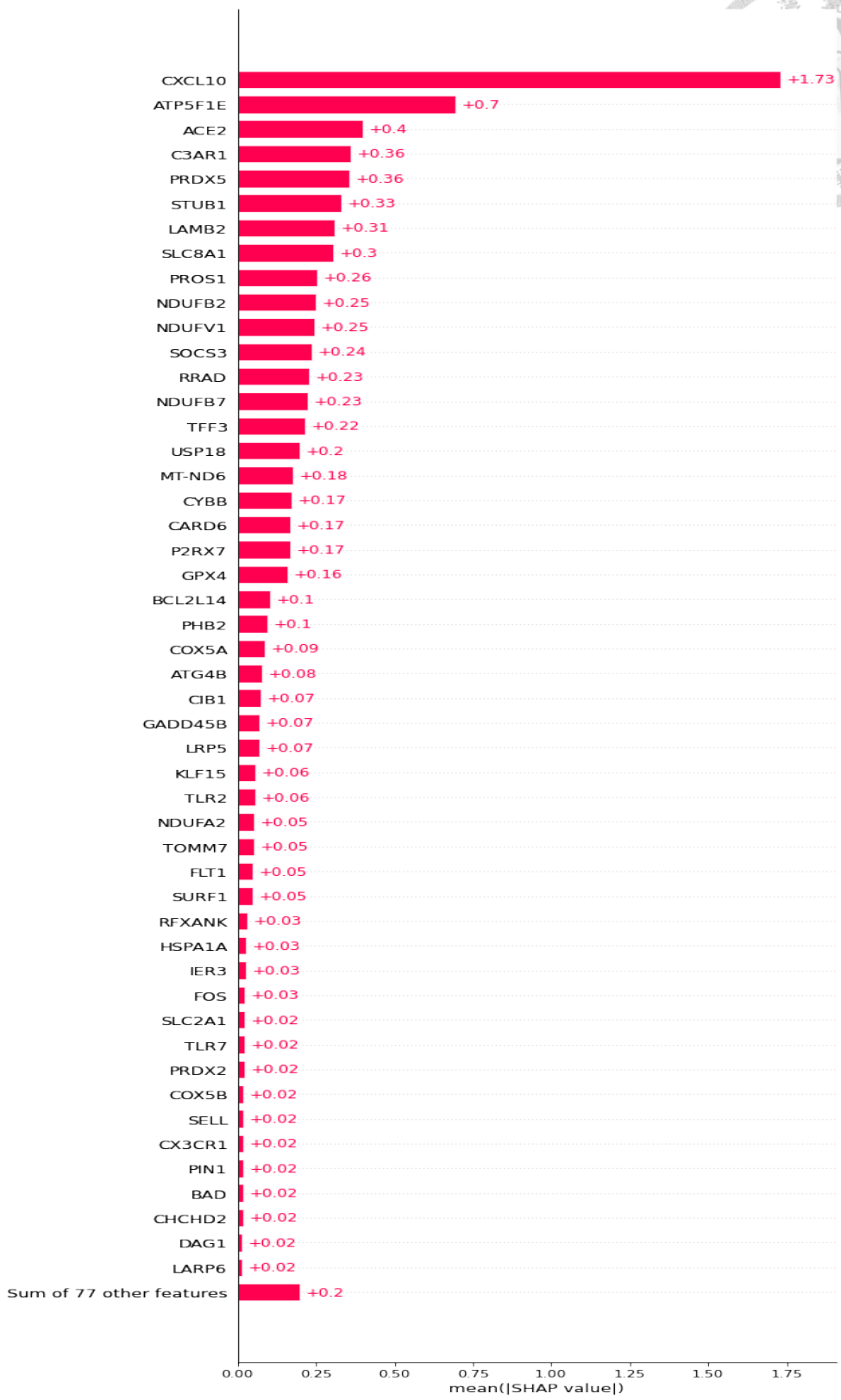
doi:10.6342/NTU202402334

acinar cells, leading to pronounced mitochondrial dysfunction and ultimately causing to apoptosis in acinar cells (68). *ATP5F1E* (ATP synthase F1 subunit epsilon) encodes a subunit of mitochondrial ATP synthase. Therefore, *ATP5F1E* can lead to defects in ATP synthase activity, resulting in a range of mitochondrial dysfunction. There has been a significant increase in the expression of ATP5F1E in patients with COVID-19, which could be linked to increased inflammation and the generation of ROS (69-71). *ACE2* (angiotensin converting enzyme 2) is expressed in various tissues, including the lungs, heart, kidneys, and intestines. It plays a significant role in protecting organs from the harmful effects of angiotensin II, such as inflammation, fibrosis, and oxidative stress. *ACE2* gained widespread recognition as the entry receptor for the SARS-CoV-2 virus. The spike protein of the virus binds to the ACE2 receptor on the surface of human cells, facilitating viral entry and infection. This interaction is a critical step in the pathogenesis of COVID-19. The expression of *ACE2* in respiratory tract cells makes it a key factor in the susceptibility and severity of COVID-19. It is also associated with mtDNA depletion and mitochondrial dysfunction (72). Mitochondria serve as the primary source of ROS in cells. Impairment of mitochondrial activity results in the excess production of ROS, which causes oxidative stress and damage to cells. *ACE2* mitigates this process, possesses protective properties, and lowers ROS levels. This helps in protecting mitochondria from oxidative damage. Besides, mitochondria also play a crucial role in cellular metabolism as they are responsible for the synthesis of ATP. Reducing *ACE2* expression or impairing its activity can lead to metabolic imbalances, which can negatively impact mitochondrial function and worsen metabolic disorders (73). Chronic inflammatory responses can be initiated by mitochondrial dysfunction, and *ACE2* plays a vital role in mitigating inflammation. *ACE2* has the potential to mitigate inflammation, thereby leading to a decrease in inflammation-induced mitochondrial damage (74). The

doi:10.6342/NTU202402334

association between *ACE2* and mitochondrial dysfunction is closely linked, specifically in the regulation of oxidative stress, cellular metabolism, and inflammatory responses. The ClueGO pathway analysis network includes selecting the most important features from genes that were significantly differentially expressed, based on their SHAP values above the average score. *CXCL10, ATP5F1E, ACE2, C3AR1, STUB1, LAMB2, SLC8A1, PROS1, SOCS3, RRAD, NDUFB7, TFF3, USP18, MT-ND6, CYBB, PHB2, COX5A, CIB1, KLF15, FLT1, BAD, NDUFV1, ATP5ME, NDUFA13, UQCRQ,* and *NDUFAB1* was selected. COVID-19 and mitochondria-related pathways were shown from the pathway analysis accordingly (Fig 9b). Thus, the top-ranking feature importance significantly expressed genes from heart-, kidney-, and liver, mitochondria-related toxicity lists were identified by a biological meaning approach reflect the relationship between COVID-19, multiorgan damage and mitochondria dysfunction and also corresponds to the predictive power of machine learning.

doi:10.6342/NTU202402334



a.

doi:10.6342/NTU202402334

b.

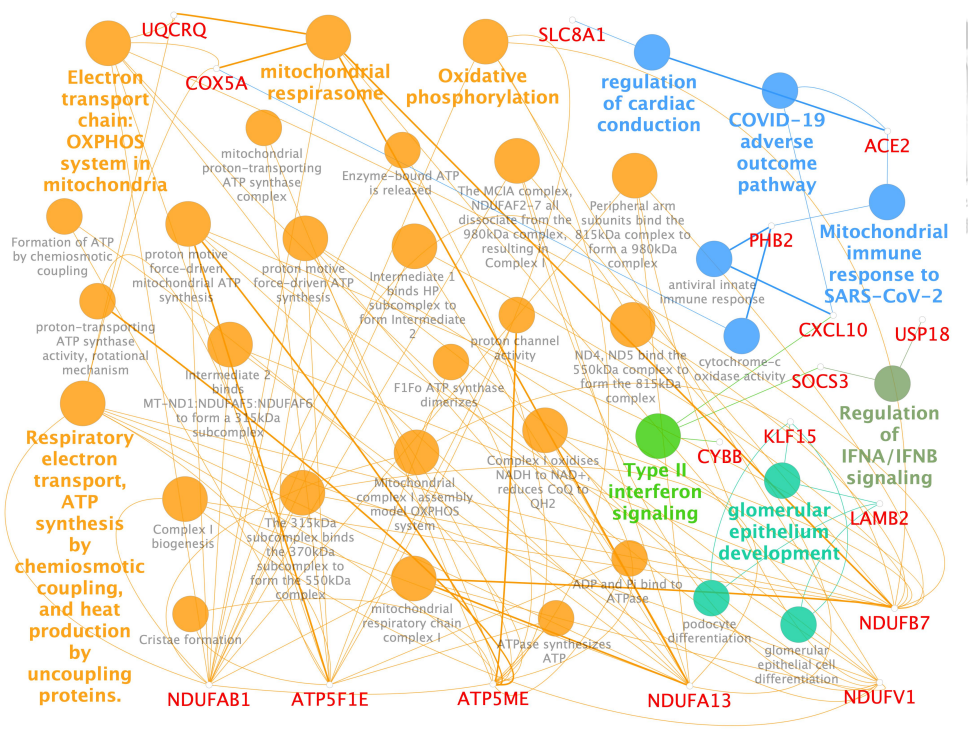


Fig 9. Feature importance analysis.

(a) Feature importance analysis from SHAP. (b) ClueGO pathway analysis network of the top ranking genes from the result of SHAP. Reuse and modified from ref. (1).

doi:10.6342/NTU202402334

Chapter 4 Discussion

The discussion will cover several key areas, including feature selection with different approaches and the use of machine learning as a tool for validation. The analysis of multiorgan damage in COVID-19 will be covered, comparing it with the multiorgan damage caused by sepsis. The role of mitochondria in the multiorgan damage induced by COVID-19 will be examined, along with the phenomena of long COVID.

4.1 Feature selection with different approaches

The utilization of machine learning has been extensively employed in the identification and assessment of patients with COVID-19. For instance, the GSE152075 sample set has been utilized in previous machine learning investigations that employed automated ML (AutoML) (75) and XGBoost for the purpose of feature selection. In the XGBoost research for feature selection, there are 24 genes which are *IGFBP2, KRT8, RPLP0, XAF1, RPL13, OAS2, CES1, RPL4, EEF1G, NR2F6, RPS8, RPL10A, SNX14, C5orf15, TNFRSF19, CD24, ALAS1, CEP112, C9orf24, POLR2J3, AAMP, DUOX2, EMCN*, and *RPL3* were selected as features for machine learning. The results showed that using KNN as the classifier achieved the highest Matthews Correlation Coefficient (MCC) value, with MCC at 0.886, sensitivity at 0.986, specificity at 0.907, and accuracy at 0.977 (76). These 24 genes selected under specific conditions that share common genes with the top 100 significant genes of GSE152075, including *RPL10A, C9orf24, OAS2, RPLP0, RPL3, XAF1, RPS8, IGFBP2*, and *RPL13*. The only common gene found among all significant genes related to mitochondria, heart, kidney, and liver in the toxicity list from GSE152075 is *KRT8*.

doi:10.6342/NTU202402334

The least absolute shrinkage and selection operator (LASSO) regression model for feature selection was also be utilized, and nasopharyngeal swab sample sets from GSE163151, GSE152075, GSE156063, and GSE188678 were applied in the research (77). Under specific conditions, there are 23 significant genes were identified including *IFI6, IFI44L, SIGLEC1, NUCB1, XAF1, TMED9, SAMHD1, SDC1, TIMM13, IL1R2, CXCL11, LAMB3, TMA7, ADIRF, BBS10, OR1I1, MIF, CXCL10, C19orf33, COPA, ADAM17, TCTEX1D4*, and *IFIT2*. Feature selection through a LASSO regression model narrowed these down to 3 to 9 genes to be used as features. Subsequently, these were used in Random Forest classification for training and prediction. The best results were obtained with eight genes: *COPA, CXCL11, IFI6, MIF, NUCB1, SAMHD1, SIGLEC1*, and *TMED9*, achieving a sensitivity of 0.872, specificity of 0.913, accuracy of 0.882, and an AUC of 0.950. These 23 significant genes overlap with the top 100 significant genes from GSE152075, including *XAF1, CXCL10, IFI44L, IFI6, CXCL11, SIGLEC1*, and *IFIT2*. Among the genes related to mitochondria, heart, kidney, and liver in GSE152075's toxicity list, the common genes include *CXCL10, ALDH3A1, NDUFV1, TNFSF13B, RRAD, PRDX5, GBP5, CIB1, CCR1, CYBB*, and *SLC8A1*. Eight genes selected through the LASSO regression model's feature selection that overlaps with GSE152075's top 100 significant genes are *IFI6, CXCL11*, and *SIGLEC1*. *MIF* is the only common gene among all significant genes related to mitochondria, heart, kidney, and liver in the toxicity list from GSE152075. Therefore, different feature selection methods result in predominantly different genes. While their predictive outcomes may be similarly effective, the biological pathways or physiological implications they represent can also differ significantly.

doi:10.6342/NTU202402334

4.2 Machine learning as the tool of validation

Typically, machine learning is employed for prediction and classification tasks. However, in this instance, it was used inversely which is to first generate a hypothesis and then to select data that conforms to this hypothesis for further machine learning analysis. If the predictive performance is robust, this strongly suggests that the hypothesis is valid, based on logical inference. Such an approach enables a meaningful interpretation of the machine learning outcomes.

The primary purpose of this study was to explore the impact of COVID-19 on mitochondrial damage and further cause the damage of multiorgan. Two aims were hypothesized and tested to establish this correlation and ascertain their validity. The first aim is to utilize a statistical approach to generate predictive models through machine learning, serving as a baseline for further analysis. This method facilitated an evaluation of the reliability and coherence of machine learning applications on the sample sets. The second aim employed a biological perspective to examine and clarify the principal assertion of this research: that COVID-19 induces mitochondrial damage, leading to subsequent damage in organs. Consequently, this investigation leveraged machine learning techniques to rigorously assess these hypotheses.

The first objective employed a statistical approach to identify significant genes that might be utilized as features in machine learning. The machine learning analysis demonstrated that prioritizing the most statistically significant genes with a differential expression made it feasible to predict COVID-19 accurately. However, this methodology did not directly reveal the impact of COVID-19 on mitochondria, possibly due to other factors exerting more direct effects. As an illustration, the examination of the top 40 genes that were significantly expressed differently showed that a substantial

doi:10.6342/NTU202402334

number of them were related to inflammatory cytokine storms. COVID-19 is strongly associated with the occurrence of inflammatory cytokine storms, which result in damage to multiple organs (78). Machine learning can accurately forecast COVID-19 by utilizing the top 40 genes that exhibit significant variations in expression as input features.

The second objective evaluated the machine learning features chosen based on biological meaning approach. It was noted that individual genes selected based on biological meaning approach may not always accurately reflect their importance, especially if they are chosen solely based on gene expression levels. For instance, only a few genes from the selected toxicity lists were among the top 100 significantly differentially expressed genes. Machine learning was employed to verify if the genes identified as significantly differentially expressed by the biological meaning approach can accurately predict COVID-19. Additionally, it was used to deduce that the damage caused by COVID-19 to mitochondria may result in subsequent harm to other organs. The conclusive study findings indicated a correlation between the influence of COVID-19 on mitochondria and subsequent cardiac, renal, and liver damage. Thus, COVID-19 might potentially harm multiple organs by causing damage to mitochondria. In this study, the analysis of toxicity on GSE152075 sample data was used to identify significantly differentially expressed genes in the heart-, liver-, kidney-, and mitochondria-related toxicity lists.

4.3 Multiorgan damage analysis in COVID-19

Significant genes identified through statistical meaning methods share few common genes with selected genes found through biological meaning approaches. However, the

doi:10.6342/NTU202402334

results from machine learning effectively predict COVID-19 infection status, albeit representing different interpretations. In this study, the features identified from the biological meaning approach are associated with mitochondrial, heart, kidney, and liver functions. Therefore, the machine learning results interpretation suggests that COVID-19 enters cells through ACE2 and directly or indirectly influences inflammatory cytokine storms, oxidative stress, and mitochondrial functions, subsequently impairing heart, liver, and kidney functions. On the other hand, the machine learning outcomes derived from features identified through statistical meaning methods suggest that one possible explanation could be that COVID-19 causes inflammatory cytokine storms, which then damage heart, liver, and kidney functions

There was little overlap between the significantly differentially expressed genes identified through the statistical meaning method and those identified through the biological meaning approach in the study. Nevertheless, both machine learning outcomes demonstrated predictive capabilities for COVID-19, with the sole discrepancy being their interpretation. The statistical approach to feature selection revealed that SARS-CoV-2 induces an inflammatory cytokine storm, leading to impaired cardiac, hepatic, and renal function. The machine learning results, analyzed from a biological perspective, revealed a correlation between SARS-CoV-2's impact on ACE2 and mitochondria and the occurrence of inflammatory cytokine storms and oxidative stress. This correlation further suggests that these effects may contribute to worsening heart, liver, and kidney conditions.

doi:10.6342/NTU202402334

4.4 The differences of multiorgan damage caused by sepsis and COVID-19

In sepsis, patients can be categorized into five distinct endotypes: Neutrophilic-Suppressive (NPS), Inflammatory (INF), Innate-Host-Defense (IHD), Interferon (IFN), and Adaptive (ADA). Each of these endotypes involves different immune responses and pathophysiological mechanisms. Pathway analysis of multiorgan damage in sepsis in the study highlights several vital pathways. Hormone secretion and interferon-related pathways indicate the significant role of the endocrine system and interferons in sepsis, particularly in immune responses against viral infections. The degranulation process of neutrophils, which release various enzymes and toxins, effectively destroys microbes but may also cause damage to host tissues.

In contrast to COVID-19, the pathway analysis of sepsis did not show mitochondria-related pathways, suggesting that mitochondrial dysfunction, a condition where the mitochondria are unable to produce enough energy for the cell, may not be a major factor in multiorgan damage in sepsis.

In COVID-19 patients, pathway analysis of multiorgan damage reveals different vital pathways. Interferon signaling and interferon alpha/beta/gamma signaling are closely related to immune responses following viral infections, underscoring the importance of interferon signaling in the antiviral process. Immune response and SARS-CoV-2-related pathways indicate that COVID-19 directly affects the immune system, triggering a strong immune response.

Additionally, the cytokine storm induced by COVID-19 is fatal, leading to extensive tissue damage and multiorgan failure. Unlike sepsis, the pathway analysis of COVID-19

doi:10.6342/NTU202402334

shows that mitochondrial dysfunction and oxidative stress play a significant role in multiorgan damage.

4.5 The role of mitochondria in multiorgan damage caused by COVID-19

The gene *BAD* is the sole gene that appears in all the toxicity lists linked to the heart, liver, kidney, and mitochondria. *BAD* regulates apoptosis, and its mRNA expression has been detected in various tissues, including the heart, liver, spleen, lung, kidney, hypothalamus, pituitary, uterine, and ovary. (79). The extrinsic death receptor pathway and the intrinsic intracellular mechanism both will initiate apoptosis and ultimately result in mitochondrial dysfunction. Effective apoptosis in hepatocytes necessitates mitochondrial dysfunction as a prerequisite for the confluence of various cell death pathways. Interactions among the Bcl-2 protein family members control the mitochondrial pathways of cell death (80). Analysis of liver biopsies obtained from patients with SARS has revealed that SARS-CoV may induce apoptosis of liver cells, resulting in liver damage (81). Apoptosis is also responsible for the decline in cardiac contractility reported in patients with COVID-19 (82). There is substantial evidence indicating that COVID-19 leads to damage in the renal tubules as a result of mitochondrial damage and apoptosis (83). Hence, mitochondrial dysfunction plays a crucial role in the process of apoptotic cell death, leading to heart, renal, and liver damage. Additionally, COVID-19 is a significant contributor to mitochondrial dysfunction.

The analysis of these common genes also revealed that SARS-CoV-2 has the potential to exploit mechanisms, such as oxidative stress, following infection of the

doi:10.6342/NTU202402334

mitochondria. Mitochondria serve as the primary origin of free radicals, which are accountable for the occurrence of oxidative stress. Oxidative stress arises when the antioxidant system cannot promptly counteract these free radicals, leading to harm inflicted upon cells and tissues. Oxidative stress explicitly affects mitochondria and their DNA due to the vulnerability of both the membrane structure and inner components of mitochondria to oxidative damage. Oxidative stress-induced damage to mitochondria directly impacts cellular energy production and metabolism, influencing all biological functions of cells and tissues. This stress induces apoptosis, subsequently linked to impaired cardiac, hepatic, and renal function.

The investigation uncovered a link between the impact of COVID-19 on mitochondria and the worsening of cardiac, hepatic, and renal function. The role of mitochondrial dysfunction in COVID-19 was identified as significant (84). Furthermore, the investigation of gene expression in A549 and Calu3 cell lines infected with SARS-CoV-2 revealed increased cytokine and inflammatory activities. Additionally, disruptions in the expression of genes related to inflammatory, mitochondrial, and autophagic processes were specifically observed in the SARS-CoV-2-infected cells (85). Mitochondria contribute to cardiac dysfunction and myocyte injury through the reduction of metabolic capacity and the generation and release of viral factors (86). This study analyzed the damage to mitochondria by analyzing the impact of SARS-CoV-2 infection on the heart, kidney, and liver. Machine learning was employed to determine if the genes associated with mitochondria, which showed significant differences in expression, could be used to predict COVID-19. Additionally, the study investigated the extent of damage to the heart, kidney, and liver resulting from mitochondrial dysfunction.

doi:10.6342/NTU202402334

4.6 Long COVID

COVID-19 and other non-COVID-19 acute respiratory infections often present with similar symptoms, but COVID-19 tends to lead to more severe illness and a higher incidence of long COVID, a condition characterized by persistent symptoms lasting for weeks or months after the acute phase of the illness. This phenomenon may be attributed to mitochondrial toxicity and oxidative stress pathways.

The analysis in this study suggests that COVID-19 may cause more severe long-term effects due to various factors:

Mitochondrial Dysfunction: Mitochondria are responsible for producing energy within cells. When viruses, particularly COVID-19, impair mitochondrial function, it can result in energy metabolism issues and cell death. The damage to mitochondria may persist even after the acute infection has resolved, leading to lingering long-term symptoms.

Oxidative Stress: COVID-19 infection can lead to excessive oxidative stress, which can cause further damage to cells and tissues. Chronic oxidative stress is associated with persistent inflammation and other long-term symptoms.

Overactive Immune Response: COVID-19 triggers a robust immune response, sometimes escalating and leading to a cytokine storm. This uncontrolled immune reaction causes severe acute symptoms and can result in ongoing immune system dysregulation and long-term symptoms.

Systemic Effects: COVID-19 affects the respiratory system and other organs, such as the cardiovascular system, liver, and kidneys. These systemic effects can explain why some individuals experience long-term symptoms affecting multiple organs.

doi:10.6342/NTU202402334

Viral Persistence: The COVID-19 virus or its fragments can persist in the body for an extended period in some individuals, continuously stimulating the immune system and contributing to persistent symptoms.

Additional research has demonstrated that COVID-19 triggers widespread host reactions and alterations in gene expression, leading to disturbances in the biological processes and functions of every organ system (87). Furthermore, research has demonstrated that patients who have recovered from COVID-19 exhibit indications of long COVID, which includes multiorgan impairment (88-90). Hence, it is imperative to conduct additional research to discover if the mitochondrial impacts of COVID-19 contribute to the eventual development of long COVID symptoms in patients (91).

doi:10.6342/NTU202402334

Chapter 5 Summary

The link between SARS-CoV-2 infection, mitochondrial dysfunction, and the subsequent development of cardiovascular disease has also been demonstrated to be crucial (92). A number of other organs, including the liver (93) and the kidneys (94), have also been found to exhibit similar findings. SARS-CoV-2, which causes COVID-19, gains entry into cells via the ACE2 receptor. Viral infection induces cellular stress and inflammation, which may directly and indirectly affect mitochondrial function, leading to mitochondrial stress and dysfunction. These effects are likely interconnected, as evidenced by the correlation between direct infection via ACE2-dependent pathways and mitochondrial dysfunction, which can result in extensive multiorgan damage (95). Thus, the multi-system inflammation and organ damage observed in COVID-19 patients could be partially attributed to mitochondrial dysfunction. Additionally, SARS-CoV-2's direct invasion of cells in various organs through ACE2 may initiate a cytokine storm, further impairing cardiac, hepatic, renal, and other vital organ functions (Fig 10).

doi:10.6342/NTU202402334

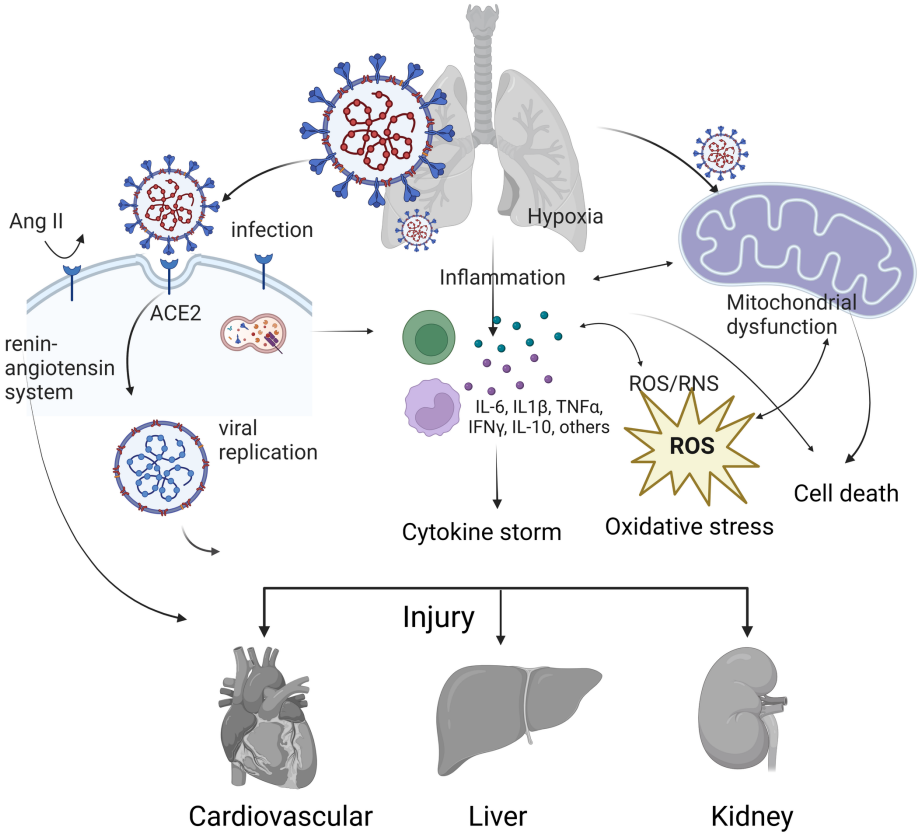


Fig 10. Multiorgan damage induced by COVID-19.

SARS-CoV-2 invades cells in various organs through ACE2, and causes cytokine storms, oxidative stress, mitochondrial dysfunction, and cell death. These effects can subsequently impair the functioning of the heart, liver, kidneys, and other organs. Reuse and modified from ref. (1).

5.1 Conclusion

This research identified genes that were expressed significantly differently in the toxicity lists of heart, liver, kidney, and mitochondria from tox analysis. The purpose was to validate their connection with COVID-19 and to conclude that COVID-19-

doi:10.6342/NTU202402334

induced mitochondrial damage exacerbates cardiac, hepatic, and renal function. Furthermore, previous research has provided evidence for the association between genes and pathways in the mitochondria, heart, liver, and kidneys with COVID-19. The objective of this study was to get the same conclusion by employing alternative methodologies that expanded the investigation and offered further interpretation for these findings. The study hypothesized that the impact of COVID-19 on mitochondria is correlated with cardiac, renal, and hepatic damage. This correlation was tested using machine learning. The study also found that SARS-CoV-2 can cause multiorgan damage through various mechanisms, such as direct invasion of cells via ACE2 and cytokine storms, impairing cardiac, hepatic, and renal function.

In sepsis, multiorgan damage is primarily caused by dysregulated immune responses and neutrophil degranulation, with less significant impact from mitochondrial dysfunction. In contrast, in COVID-19, multiorgan damage is driven by multiple mechanisms, including interferon signaling, cytokine storms, oxidative stress, and mitochondrial dysfunction.

These differences highlight the distinct mechanisms of multiorgan damage in sepsis and COVID-19, emphasizing the need for different therapeutic strategies for these two conditions.

There is another finding suggests that there are similarities when comparing the pathways and toxicity analyses among patients with COVID-19, other non-COVID-19 acute respiratory infections, and healthy individuals. However, a more detailed comparison between COVID-19 and other non-COVID-19 acute respiratory infections reveals differences in pathways linked to mitochondrial toxicity and oxidative stress. These distinctions may explain why COVID-19 often leads to more severe symptoms and a higher incidence of long COVID. Despite the similarities in symptoms between

doi:10.6342/NTU202402334

COVID-19 and other non-COVID-19 acute respiratory infections, COVID-19 generally results in more severe illness and a higher prevalence of long COVID.

In summary, the long-term symptoms caused by COVID-19 may stem from a combination of factors, including mitochondrial dysfunction, oxidative stress, overactive immune response, and multi-organ system effects. These interrelated factors may account for the increased incidence and severity of long COVID compared to other non-COVID-19 acute respiratory infections.

The results of toxicity lists exhibited variability among the different examined tissue samples, and the results from the same tissues could also vary depending on the sampling platform and the ratio or size of the experimental and control samples. Subsequent investigation may find value in further interpreting and utilizing machine learning to examine the impact of various sample tissues and sizes on the hypothesis.

5.2 Limitations and potential problems

The performance of machine learning models trained on GSE152075 data sets is significantly poor when evaluated on other sample sets. Consequently, the learned machine learning model cannot be effectively applied to diverse sample sets, whether there are variations in the tissues, gene expression detection platforms, or parameter settings. Moreover, using sample sets containing diverse tissues may yield dissimilar outcomes compared to this study owing to tissue-specific factors.

doi:10.6342/NTU202402334

5.3 Future work

Future work will involve comparing the results of imbalanced data implemented with or without differential gene expression (DGE) analysis and machine learning. Additionally, the focus will be on optimizing strategies for handling imbalanced data in DGE analysis and machine learning. Furthermore, the research will aim to identify the optimal machine learning algorithm and strategy tailored to different sample data statuses, tissue types, and sample sizes.

References

1. Chang YY, Wei AC. Transcriptome and machine learning analysis of the impact of COVID-19 on mitochondria and multiorgan damage. PLoS One. 2024;19(1):e0297664.
2. Mizock BA. The multiple organ dysfunction syndrome. Dis Mon. 2009;55(8):476-526.
3. Nakayama R, Bunya N, Tagami T, Hayakawa M, Yamakawa K, Endo A, et al. Associated organs and system with COVID-19 death with information of organ support: a multicenter observational study. BMC Infect Dis. 2023;23(1):814.
4. Thakur V, Ratho RK, Kumar P, Bhatia SK, Bora I, Mohi GK, et al. Multi-Organ Involvement in COVID-19: Beyond Pulmonary Manifestations. J Clin Med. 2021;10(3).
5. Iacobucci G. Long covid: Damage to multiple organs presents in young, low risk patients. BMJ. 2020;371:m4470.
6. Brauninger H, Stoffers B, Fitzek ADE, Meissner K, Aleshcheva G, Schweizer M, et al. Cardiac SARS-CoV-2 infection is associated with pro-inflammatory transcriptomic alterations within the heart. Cardiovasc Res. 2022;118(2):542-55.
7. Uribarri A, Nunez-Gil IJ, Aparisi A, Becerra-Munoz VM, Feltes G, Trabattoni D, et al. Impact of renal function on admission in COVID-19 patients: an analysis of the international HOPE COVID-19 (Health Outcome Predictive Evaluation for COVID 19) Registry. J Nephrol. 2020;33(4):737-45.
8. Saha L, Vij S, Rawat K. Liver injury induced by COVID 19 treatment - what do we know? World J Gastroenterol. 2022;28(45):6314-27.
9. Gonzalez MA, Ochoa CD. Multiorgan System Failure in Sepsis. Sepsis2018. p. 67-71.
10. Gustot T. Multiple organ failure in sepsis: prognosis and role of systemic inflammatory response. Curr Opin Crit Care. 2011;17(2):153-9.
11. Martin Gimenez VM, de Las Heras N, Ferder L, Lahera V, Reiter RJ, Manucha W. Potential Effects of Melatonin and Micronutrients on Mitochondrial Dysfunction during a Cytokine Storm Typical of Oxidative/Inflammatory Diseases. Diseases. 2021;9(2).
12. Kozlov AV, Lancaster JR, Jr., Meszaros AT, Weidinger A. Mitochondria-meditated pathways of organ failure upon inflammation. Redox Biol. 2017;13:170-81.
13. Ganji R, Reddy PH. Impact of COVID-19 on Mitochondrial-Based Immunity in Aging and Age-Related Diseases. Front Aging Neurosci. 2020;12:614650.

doi:10.6342/NTU202402334

14. Iwasaki M, Saito J, Zhao H, Sakamoto A, Hirota K, Ma D. Inflammation Triggered by SARS-CoV-2 and ACE2 Augment Drives Multiple Organ Failure of Severe COVID-19: Molecular Mechanisms and Implications. Inflammation. 2021;44(1):13-34.
15. Srinivasan K, Pandey AK, Livingston A, Venkatesh S. Roles of host mitochondria in the development of COVID-19 pathology: Could mitochondria be a potential therapeutic target? Mol Biomed. 2021;2:38.
16. Guarnieri JW, Dybas JM, Fazelinia H, Kim MS, Frere J, Zhang Y, et al. Targeted Down Regulation Of Core Mitochondrial Genes During SARS-CoV-2 Infection. bioRxiv. 2022.
17. Vivaldi G, Pfeffer PE, Talaei M, Basera TJ, Shaheen SO, Martineau AR. Long-term symptom profiles after COVID-19 vs other acute respiratory infections: an analysis of data from the COVIDENCE UK study. EClinicalMedicine. 2023;65:102251.
18. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019;47(W1):W234-W41.
19. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216-W21.
20. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
21. Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523-30.
22. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091-3.
23. 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(11):2498-504.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.
25. Baghela A, Pena OM, Lee AH, Baquir B, Falsafi R, An A, et al. Predicting sepsis severity at first clinical presentation: The role of endotypes and mechanistic signatures. EBioMedicine. 2022;75:103776.
26. Ng DL GA, Santos YA, Servellita V et al. A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood. Sci Adv 2021 Feb(eabe5984).
27. Lieberman NAP, Peddu V, Xie H, Shrestha L, Huang ML, Mears MC, et al. In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. PLoS Biol. 2020;18(9):e3000849.
28. Overmyer KA, Shishkova E, Miller IJ, Balnis J, Bernstein MN, Peters-Clarke TM, et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst. 2021;12(1):23-40 e7.
29. Thair SA, He YD, Hasin-Brumshtein Y, Sakaram S, Pandya R, Toh J, et al. Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections. iScience. 2021;24(1):101947.
30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
31. Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values? Cell J. 2019;20(4):604-7.
32. Zhao B, Erwin A, Xue B. How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances. Genomics. 2018;110(1):67-73.
33. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics. 2015;16(1):169.
34. Goedhart J, Luijsterburg MS. VolcaNoseR is a web app for creating, exploring, labeling and sharing volcano plots. Sci Rep. 2020;10(1):20560.

doi:10.6342/NTU202402334

35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
36. Chen TQG, C. XGBoost- A Scalable Tree Boosting System. arXiv:160302754v3. 2016.
37. Breiman L. Random Forests. Machine Learning, 45, 5-32. 2001.
38. Boateng EY, Abaye DA. A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing. 2019;07(04):190-207.
39. Noble WS. What is a support vector machine. Nature Biotechnology. 2006;24:pages1565–7.
40. Hicks SA, Strumke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1):5979.
41. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Kaunas). 2021;57(5).
42. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021;14(1):13.
43. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak. 2020;20(1):4.
44. Marcilio WE, Eler DM. From explanations to feature selection: assessing SHAP values as feature selection mechanism. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)2020. p. 340-7.
45. Gupta T, Puskarich MA, DeVos E, Javed A, Smotherman C, Sterling SA, et al. Sequential Organ Failure Assessment Component Score Prediction of In-hospital Mortality From Sepsis. J Intensive Care Med. 2020;35(8):810-7.
46. Baghela A, An A, Zhang P, Acton E, Gauthier J, Brunet-Ratnasingham E, et al. Predicting severity in COVID-19 disease using sepsis blood gene expression signatures. Sci Rep. 2023;13(1):1247.
47. Ramasamy S, Subbian S. Critical Determinants of Cytokine Storm and Type I Interferon Response in COVID-19 Pathogenesis. Clin Microbiol Rev. 2021;34(3).
48. Streng L, de Wijs CJ, Raat NJH, Specht PAC, Sneiders D, van der Kaaij M, et al. In Vivo and Ex Vivo Mitochondrial Function in COVID-19 Patients on the Intensive Care Unit. Biomedicines. 2022;10(7).
49. Wajihah Mughal LAK. Cell death signalling mechanisms in heart failure. Exp Clin Cardiol. 2011 Winter:16(4): 102–8.
50. Jiang W, Xiong Y, Li X, Yang Y. Cardiac Fibrosis: Cellular Effectors, Molecular Pathways, and Exosomal Roles. Front Cardiovasc Med. 2021;8:715258.
51. Guicciardi ME, Malhi H, Mott JL, Gores GJ. Apoptosis and necrosis in the liver. Compr Physiol. 2013;3(2):977-1010.
52. Wanless IR, Lentz JS. Fatty liver hepatitis (steatohepatitis) and obesity: an autopsy study with analysis of risk factors. Hepatology. 1990;12(5):1106-10.
53. MARIE-REINE LOSSER DP. Mechanisms of Liver Damage. SEMINARS IN LIVER DISEASE. 1996;16.
54. Brenner BM. Hemodynamically mediated glomerular injury and the progressive nature of kidney disease. Kidney Int. 1983;23(4):647-55.
55. Salant DJ, Quigg RJ, Cybulsky AV. Heymann nephritis: mechanisms of renal injury. Kidney Int. 1989;35(4):976-84.
56. Colombo M, Valo E, McGurnaghan SJ, Sandholm N, Blackbourn LAK, Dalton RN, et al. Biomarker panels associated with progression of renal disease in type 1 diabetes. Diabetologia. 2019;62(9):1616-27.
57. Priante G, Gianesello L, Ceol M, Del Prete D, Anglani F. Cell Death in the Kidney. Int J Mol Sci. 2019;20(14).
58. Lei X, Du L, Yu W, Wang Y, Ma N, Qu B. GSTP1 as a novel target in radiation induced lung injury. J Transl Med. 2021;19(1):297.

doi:10.6342/NTU202402334

59. NCBI. BAD BCL2 associated agonist of cell death [Homo sapiens (Human)] - Gene. https://wwwncbinlmnihgov/gene?Db=gene&Cmd=DetailsSearch&Term=572. 2023.
60. Khan AUH, Rathore MG, Allende-Vega N, Vo DN, Belkhala S, Orecchioni S, et al. Human Leukemic Cells performing Oxidative Phosphorylation (OXPHOS) Generate an Antioxidant Response Independently of Reactive Oxygen species (ROS) Production. EBioMedicine. 2016;3:43-53.
61. Wang CH, Wu SB, Wu YT, Wei YH. Oxidative stress response elicited by mitochondrial dysfunction: implication in the pathophysiology of aging. Exp Biol Med (Maywood). 2013;238(5):450-60.
62. Sinha K, Das J, Pal PB, Sil PC. Oxidative stress: the mitochondria-dependent and mitochondria-independent pathways of apoptosis. Arch Toxicol. 2013;87(7):1157-80.
63. Samaras C, Kyriazopoulou E, Poulakou G, Reiner E, Kosmidou M, Karanika I, et al. Interferon gamma-induced protein 10 (IP-10) for the early prognosis of the risk for severe respiratory failure and death in COVID-19 pneumonia. Cytokine. 2023;162:156111.
64. Coperchini F, Chiovato L, Rotondi M. Interleukin-6, CXCL10 and Infiltrating Macrophages in COVID-19-Related Cytokine Storm: Not One for All But All for One! Front Immunol. 2021;12:668507.
65. N. ZHANG Y-DZ, X.-M. WANG. CXCL10 an important chemokine associated with cytokine storm in COVID-19 infected patients. Eur Rev Med Pharmacol Sci. 2020 Jul.(24(13):7497-7505.).
66. Gudowska-Sawczuk M, Mroczko B. What Is Currently Known about the Role of CXCL10 in SARS-CoV-2 Infection? Int J Mol Sci. 2022;23(7).
67. Lore NI, De Lorenzo R, Rancoita PMV, Cugnata F, Agresti A, Benedetti F, et al. CXCL10 levels at hospital admission predict COVID-19 outcome: hierarchical assessment of 53 putative inflammatory biomarkers in an observational study. Mol Med. 2021;27(1):129.
68. Singh L, Arora SK, Bakshi DK, Majumdar S, Wig JD. Potential role of CXCL10 in the induction of cell injury and mitochondrial dysfunction. Int J Exp Pathol. 2010;91(3):210-23.
69. He J, Cai S, Feng H, Cai B, Lin L, Mai Y, et al. Single-cell analysis reveals bronchoalveolar epithelial dysfunction in COVID-19 patients. Protein Cell. 2020;11(9):680-7.
70. Santos AF, Povoa P, Paixao P, Mendonca A, Taborda-Barata L. Changes in Glycolytic Pathway in SARS-COV 2 Infection and Their Importance in Understanding the Severity of COVID-19. Front Chem. 2021;9:685196.
71. Pietrobon AJ, Andrejew R, Custodio RWA, Oliveira LM, Scholl JN, Teixeira FME, et al. Dysfunctional purinergic signaling correlates with disease severity in COVID-19 patients. Front Immunol. 2022;13:1012027.
72. Zhao Q, Zhou X, Kuiper R, Curbo S, Karlsson A. Mitochondrial dysfunction is associated with lipid metabolism disorder and upregulation of angiotensin-converting enzyme 2. PLoS One. 2022;17(6):e0270418.
73. Cao X, Song LN, Yang JK. ACE2 and energy metabolism: the connection between COVID-19 and chronic metabolic disorders. Clin Sci (Lond). 2021;135(3):535-54.
74. Dai S, Cao T, Shen H, Zong X, Gu W, Li H, et al. Landscape of molecular crosstalk between SARS-CoV-2 infection and cardiovascular diseases: emphasis on mitochondrial dysfunction and immune-inflammation. J Transl Med. 2023;21(1):915.
75. Papoutsoglou G, Karaglani M, Lagani V, Thomson N, Roe OD, Tsamardinos I, et al. Automated machine learning optimizes and accelerates predictive modeling from COVID-19 high throughput datasets. Sci Rep. 2021;11(1):15107.
76. Song X, Zhu J, Tan X, Yu W, Wang Q, Shen D, et al. XGBoost-Based Feature Learning Method for Mining COVID-19 Novel Diagnostic Markers. Front Public Health. 2022;10:926069.
77. Maleknia S, Tavassolifar MJ, Mottaghitalab F, Zali MR, Meyfour A. Identifying novel host-based diagnostic biomarker panels for COVID-19: a whole-blood/nasopharyngeal transcriptome meta-analysis. Mol Med. 2022;28(1):86.
78. Hu B, Huang S, Yin L. The cytokine storm and COVID-19. J Med Virol. 2021;93(1):250-6.

doi:10.6342/NTU202402334

79. Cao X, Wang X, Lu L, Li X, Di R, He X, et al. Expression and Functional Analysis of the BCL2-Associated Agonist of Cell Death (BAD) Gene in the Sheep Ovary During the Reproductive Cycle. Front Endocrinol (Lausanne). 2018;9:512.
80. Cazanave SC, Gores GJ. The liver's dance with death: two Bcl-2 guardian proteins from the abyss. Hepatology. 2009;50(4):1009-13.
81. Xu L, Liu J, Lu M, Yang D, Zheng X. Liver injury during highly pathogenic human coronavirus infections. Liver Int. 2020;40(5):998-1004.
82. Tangos M, Budde H, Kolijn D, Sieme M, Zhazykbayeva S, Lodi M, et al. SARS-CoV-2 infects human cardiomyocytes promoted by inflammation and oxidative stress. Int J Cardiol. 2022;362:196-205.
83. Alexander MP, Mangalaparthi KK, Madugundu AK, Moyer AM, Adam BA, Mengel M, et al. Acute Kidney Injury in Severe COVID-19 Has Similarities to Sepsis-Associated Kidney Injury: A Multi-Omics Study. Mayo Clin Proc. 2021;96(10):2561-75.
84. Moreno Fernandez-Ayala DJ, Navas P, Lopez-Lluch G. Age-related mitochondrial dysfunction as a key factor in COVID-19 disease. Exp Gerontol. 2020;142:111147.
85. Singh K, Chen YC, Hassanzadeh S, Han K, Judy JT, Seifuddin F, et al. Network Analysis and Transcriptome Profiling Identify Autophagic and Mitochondrial Dysfunctions in SARS-CoV-2 Infection. Front Genet. 2021;12:599261.
86. Lesnefsky EJ, Moghaddas S, Tandler B, Kerner J, Hoppel CL. Mitochondrial dysfunction in cardiac disease: ischemia--reperfusion, aging, and heart failure. J Mol Cell Cardiol. 2001;33(6):1065-89.
87. Park J, Foox J, Hether T, Danko DC, Warren S, Kim Y, et al. System-wide transcriptome damage and tissue identity loss in COVID-19 patients. Cell Rep Med. 2022;3(2):100522.
88. Al-Aly Z, Bowe B, Xie Y. Long COVID after breakthrough SARS-CoV-2 infection. Nat Med. 2022;28(7):1461-7.
89. Yan Z, Yang M, Lai CL. Long COVID-19 Syndrome: A Comprehensive Review of Its Effect on Various Organ Systems and Recommendation on Rehabilitation Plans. Biomedicines. 2021;9(8).
90. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol. 2023:1-14.
91. Nunn AVW, Guy GW, Brysch W, Bell JD. Understanding Long COVID; Mitochondrial Health and Adaptation-Old Pathways, New Problems. Biomedicines. 2022;10(12).
92. Chang X, Ismail NI, Rahman A, Xu D, Chan RWY, Ong SG, et al. Long COVID-19 and the Heart: Is Cardiac Mitochondria the Missing Link? Antioxid Redox Signal. 2022;38(7-9):599-618.
93. Wang X, Lei J, Li Z, Yan L. Potential Effects of Coronaviruses on the Liver: An Update. Front Med (Lausanne). 2021;8:651658.
94. Ronco C, Reis T, Husain-Syed F. Management of acute kidney injury in patients with COVID-19. Lancet Respir Med. 2020;8(7):738-42.
95. Kirtipal N, Kumar S, Dubey SK, Dwivedi VD, Gireesh Babu K, Maly P, et al. Understanding on the possible routes for SARS CoV-2 invasion via ACE2 in the host linked with multiple organs damage. Infect Genet Evol. 2022;99:105254.

Appendix

doi:10.6342/NTU202402334