RAMEN Unveils Clinical Variable Networks for COVID-19 Severity and Long COVID Using Absorbing Random Walks and Genetic Algorithms

Yiwei Xiong[1†], Jingtao Wang[1,2†], Xiaoxiao Shang[1,3], Tingting Chen[4], Douglas D. Fraser[5,6,7,8,9], Gregory Fonseca[1,2], Simon Rousseau[1,2*], Jun Ding[1,2,10,11*]

[1]Meakins-Christe Laboratories, Research Institute of McGill University Health Centre, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada.
[2]Department of Medicine, Division of Experimental Medicine, McGill University, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada.
[3]Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, H3A 0B9, Quebec, Canada.
[4]Hematology Department, Capital Medical University, Beijing, China.
[5]Children's Health Research Institute, Victoria Research Laboratories, 800 Commissioners Road East, London, N6C 2V5, Ontario, Canada.
[6]Lawson Health Research Institute, London, N6C 2R5, Ontario, Canada.
[7]Department of Pediatrics, Western University, London, N6A 5C1, Ontario, Canada.
[8]Department of Physiology & Pharmacology, Western University, London, N6A 5C1, Ontario, Canada.
[9]Department of Clinical Neurological Sciences, Western University, London, N6A 5C1, Ontario, Canada.
[10]School of Computer Science, McGill University, 3480 Rue University, Montreal, H3A 2A7, Quebec, Canada.
[11]Mila-Quebec AI Institute, 6666 Rue Saint-Urbain, Montreal, H2S 3H1, Quebec, Canada.

*Corresponding author(s). E-mail(s): simon.rousseau@mcgill.ca; jun.ding@mcgill.ca;
Contributing authors: yiwei.xiong@mail.mcgill.ca; jingtao.wang@mail.mcgill.ca; xiaoxiao.shang@mail.mcgill.ca; bjlhyyxyk@ccmu.edu.cn; douglas.fraser@lhsc.on.ca; gregory.fonseca@mcgill.ca;
†These authors contributed equally to this work.

Abstract

The COVID-19 pandemic has significantly altered global socioeconomic structures and individual lives. Understanding the disease mechanisms and facilitating diagnosis requires comprehending the complex interplay among clinical factors like demographics, symptoms, comorbidities, treatments, lab results, complications, and other metrics, and their relation to outcomes such as disease severity and long term outcomes (*e.g.*, post-COVID-19 condition/long COVID). Conventional correlational methods struggle with indirect and directional connections among these factors, while standard graphical methods like Bayesian networks are computationally demanding for extensive clinical variables. In response, we introduced RAMEN, a methodology that integrates Genetic Algorithms with random walks for efficient Bayesian network inference, designed to map the intricate relationships among clinical variables. Applying RAMEN to the Biobanque québécoise de la COVID-19 (BQC19) dataset, we identified critical markers for long COVID and varying disease severity. The Bayesian Network, corroborated by existing literature and supported through multi-omics analyses, highlights significant clinical variables linked to COVID-19 outcomes. RAMEN's ability to accurately map these connections contributes substantially to developing early and effective diagnostics for severe COVID-19 and long COVID.

Keywords: Post-COVID Conditions, Long COVID, COVID-19 Severity, Genetic Algorithm, Random Walk, Bayesian Network, Clinical Variables, Multi-omics

Introduction

The outbreak of the COVID-19 pandemic has reshaped billions of lives across the globe and caused catastrophic socioeconomic losses in the past years[1, 2]. Despite the impact that COVID-19 has brought to humanity, a comprehensive understanding of COVID-19 disease mechanisms remains elusive, which substantially limits the diagnostics and therapeutics for COVID-19. For instance, it is known that different COVID-19 patients develop distinct symptoms and clinical outcomes[3]. However, it remains unclear why some people tend to develop more severe COVID-19 outcomes while some barely show any symptoms. This lack of understanding restricts the early clinical diagnosis and interventions for the most vulnerable COVID-19 victims[4, 5], which leads to undoing suffering. Moreover, although most COVID-19 victims can recover from the disease, approximately 10-30% of them develop long-term symptoms (termed as long COVID or post-COVID-19 conditions)[4], which has a tremendous socioeconomic impact. Affected individuals endure both physical and psychological distress[5]. The loss of work capacity among many sufferers has notably hindered economic productivity. The fast-accumulating COVID-19 datasets present unprecedented opportunities to derive a deeper understanding of the disease mechanisms underlying COVID-19, which will drive the discovery of novel diagnostic markers and therapeutic targets. Multiple initiatives across the globe are profiling clinical variables and genomics sequencing associated with COVID-19 patients[6, 7]. In Quebec, Canada, the Biobanque québécoise de la COVID-19 (BQC19)[8] has collected clinical data for over six thousand COVID-19 patients, along with data from several other modalities including proteomics and transcriptomics on a substantial subset. In Ontario, Canada, investigators at the Lawson Health Research Institute have also collected COVID-19 patient clinical information data [9]. These valuable data resources provide us with the opportunity to develop novel computational methods for better understanding the disease, and coming up with potential diagnosis methods and treatments,

With the ever-increasing availability of COVID-19 datasets, many studies interrogated the complex relationships between various clinical variables and COVID-19 severity levels [10–13] using simple statistical methods (e.g., correlation[13] and mutual information [14]). Existing methodologies excel at mapping direct relationships between clinical variables and outcomes, delineating these associations as edges within a network. However, while standard statistical tools like Pearson correlation and mutual information are proficient at inferring associations between variables, they do not elucidate the directionality of these relationships. These methods might identify direct connections between pairs of variables, yet they often overlook the indirect relationships crucial in complex networks composed of hundreds of variables. Moreover, statistical approaches focused on mutual information or correlation primarily examine pairwise relationships, leaving the complex interactions among multiple variables largely unaddressed. Additionally, although direct correlational analysis can effectively highlight associations between two variables based on direct correlation tests, it does not necessarily imply that these associations are relevant or predictive of the disease outcome. In essence, the relationships identified through correlation analysis may not be directly connected to the disease outcome, revealing a significant limitation in the application of such analyses for comprehensively understanding disease mechanisms and their impacts.

On the other hand, Bayesian Network (BN)[15–17], a probabilistic graphical model, can address the above limitations of simple statistical approaches, enabling the inference of clinical variables that could indicate COVID-19 outcomes (e.g., COVID-19 severity or long COVID). Previously, BNs have shown success in many application scenarios and outperformed physicians in disease diagnostics[18, 19]. For example, a BN-based diagnosis not only achieved state-of-the-art overall performance for neurodegenerative diseases compared to a list of other methods [20–25] but also provided very good interpretability. However, building a BN (particularly structure learning) for hundreds and thousands of clinical variables is challenging and computationally expensive. This task is complicated by the huge search space of possible solutions (In fact, the problem is NP-hard[26]). In practice, the BN structure learning is often regularized by prior knowledge (e.g., known constraints) to reduce the search space[27, 28]. Unfortunately, this is not an option for COVID-19 since our current understanding of this relatively new disease remains very limited[29]. Additionally, since our major aim is to find new clinical variables that influence COVID-19 outcomes (COVID-19 severity or long COVID) either directly or via another clinical variable, relying on prior knowledge unavoidably introduces bias.

To address the above limitations and fill the gap, here we introduce RAMEN (Random walk and genetic AlgorithM-based nEtwork iNference) that glues random walk and Genetic Algorithm to infer a Bayesian network representing the relationships between clinical variables in COVID-19 (particularly between clinical variables and COVID-19 outcomes). The random walk is employed to rank and select the most relevant variables and connections to COVID-19 outcomes to reduce the network complexity.

A significant aspect of our methodology is the incorporation of a terminal absorbing node, symbolizing the disease outcome of interest, such as COVID-19 severity, within our clinical variable network. In the process of conducting a random walk, all paths culminate at this terminal absorbing node. This approach effectively assigns a "direction" towards the disease outcome for the random walk, thereby facilitating the identification of network edges (associations) that have a direct link to the disease outcome (absorbing node). This innovative feature enhances the capability to discern which associations are most relevant to understanding and predicting the disease outcome. Following the preliminary network reconstruction through the absorbing random walk process, we employ a Genetic Algorithm to refine and identify an optimized network structure. This optimized structure is more accurately aligned with the observed clinical variables, ensuring a precise representation of the relationships and interactions within the dataset. After these two stages, RAMEN outputs a BN that models the complex relationship between the COVID-19 outcome variable (*e.g.*, severity and long-COVID) and other variables that are directly or indirectly connected to it. To examine the performance of RAMEN, we applied the method on three different COVID-19 cohorts from the BQC19 project and Lawson Health Research Institute, examined the resulting network with multi-omics measurements and computational simulations, and compared it with other methods. We show the resulting networks capture important COVID-19 outcome indicators that can be validated via multi-omics, simulation, or literature. Moreover, RAMEN demonstrated superior performance over simple statistical methods by finding more relevant variables and indirect variables that cannot be found using simple statistical methods such as mutual information and Pearson correlation.

Our model's ability to predict early indicators of COVID-19 outcomes, such as severity and long COVID, significantly enhances early diagnostic development for patients at risk of severe symptoms or long COVID. This advancement enables personalized and timely diagnosis, streamlining targeted treatment strategies. Such precision in diagnosis and treatment not only improves patient outcomes but also bolsters our fight against the pandemic by allowing for more efficient resource allocation and reducing the spread and impact of the virus on individuals and healthcare systems. Furthermore, this model has the potential to be generalized for analyzing clinical variable networks in diseases beyond COVID-19, provided that similar records of clinical variables are available. This enhances its utility across a wider spectrum of medical research.

Results

Overview of RAMEN

The RAMEN method comprises two phases: first, RAMEN applies absorbing random walks to select the most relevant variables to the COVID-19 outcomes (severity or long COVID) and builds a draft candidate network. Second, RAMEN employs a Genetic Algorithm to search the optimized network that represents the relationships between different clinical variables based on the draft network from the random walks (Fig. 1).

For the random walk phase 1, we first initialize a full network composing all variables as nodes, and then compute edge weights as the mutual information between the node pairs. The clinical variable of interest (often disease outcome) will be regarded as the absorbing node, for which only incoming edges are allowed. All other nodes in the graph will be regarded as non-terminal. We then perform random walks starting from each non-terminal variable (all variables other than the COVID-19 outcomes) for N random steps. At each node, the random walk uses the normalized mutual information stored in outgoing edges as transition probabilities. A random walk will terminate successfully if and only if it reaches the absorbing (terminal) node within N steps. Edge visits of successful random walks will be recorded. Intuitively, edges with more visits are the ones with stronger connections and are relevant to the COVID-19 outcome node. Subsequently, to find the edges with significantly more visits, we perform a permutation test. Specifically, we run another version of random walk on a network with permuted edge weights and use the edge visits in this network as the random background. Finally, we fit a Negative Binomial distribution to the random background (see Supplementary Fig. S1 for the histogram of edge visit count) and select edges with significantly more visits (q-value ≤ 0.05) to form the initial network skeleton.

Utilizing the Genetic Algorithm (GA), our method refines the Bayesian network structure by starting with an initial network skeleton and generating a diverse pool of initial network candidates in phase 2. These candidates are created by randomly assigning directions to the bidirectional edges in the skeleton. The evolutionary process involves performing crossover between candidate pairs to produce combined networks (offspring), introducing mutations to these offspring to create variability, and then scoring all

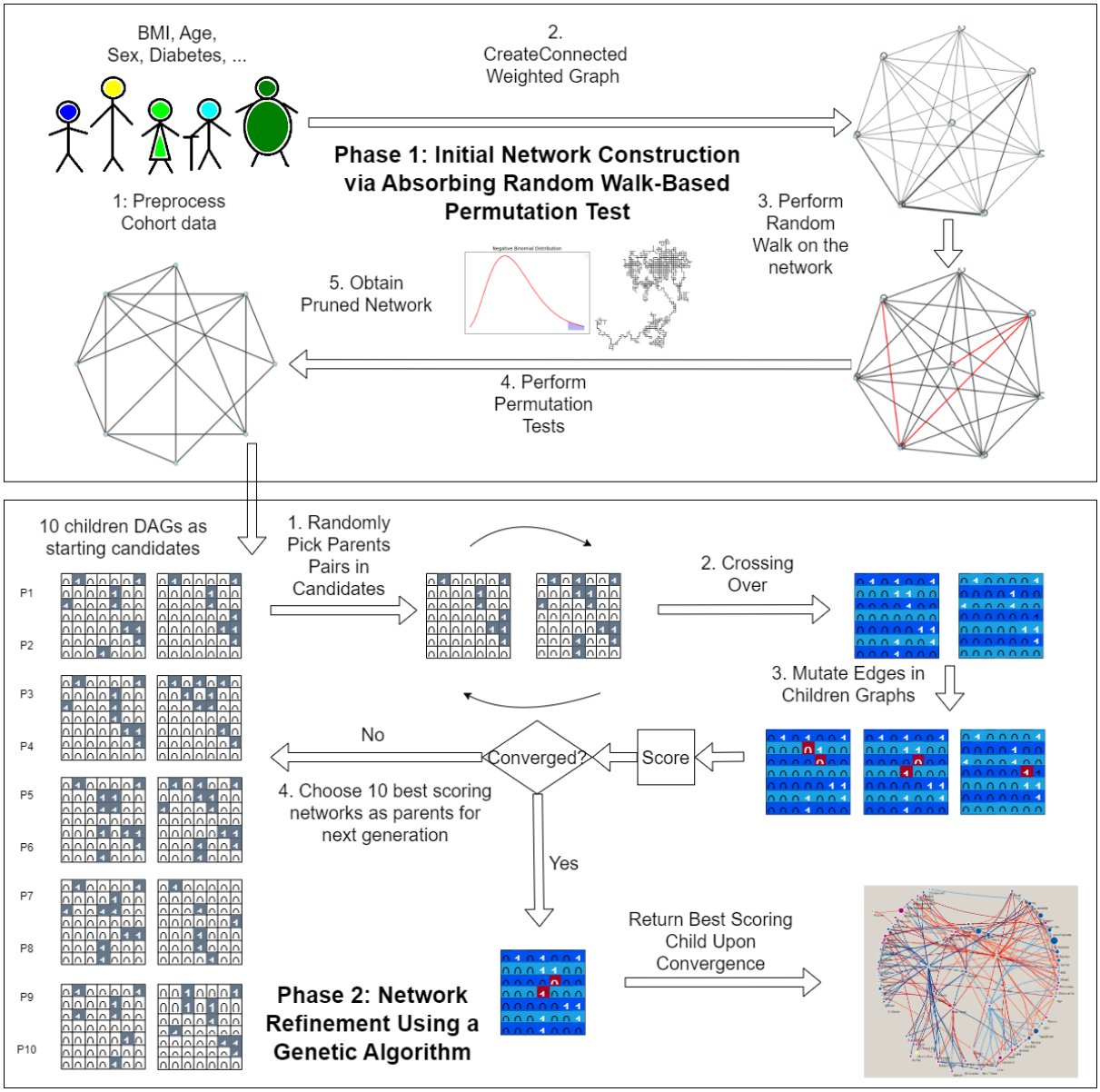

Fig. 1 Overview of the RAMEN Methodology. The RAMEN approach constructs Bayesian networks from clinical data through a sequential two-phase process. **Phase 1: Establishing the Initial Network with Absorbing Random Walk-Based Permutation Test.** Beginning with preprocessed clinical data, this stage implements a permutation test via a random walk strategy across a comprehensive network of all included variables, where nodes symbolize variables and edge weights indicate the mutual information among variable pairs. The process identifies stronger variable connections by tracking the frequency of edge traversal in successful random walks (ending at the target node). Edges with significantly higher traversal frequencies, as established through permutation testing, lay the groundwork for the network, preparing it for further enhancement. **Phase 2: Enhancing the Network with a Genetic Algorithm.** This stage further refines the Bayesian network structure. Starting with a set of initial network configurations derived from the early framework, the Genetic Algorithm applies crossover (merging two configurations) and mutation (applying random changes) to evolve these structures. Each cycle assesses the network structures against a specific scoring function, prioritizing those with superior scores for subsequent iterations. This cycle of refinement, through modification, assessment, and selection, persists until a stable score is achieved, culminating in an optimized network structure.

candidates—parents, offspring, and mutated offspring—using an Entropy-based scoring function. The highest-scoring networks are selected for the next iteration. This cycle of crossover, mutation, evaluation, and selection is iterated until the scoring converges, at which point the final network is obtained. This approach systematically explores and refines potential network structures, ensuring the development of an optimized Bayesian network that accurately models the clinical data.

Building COVID-19 severity network using the BQC19 hospitalized patient dataset

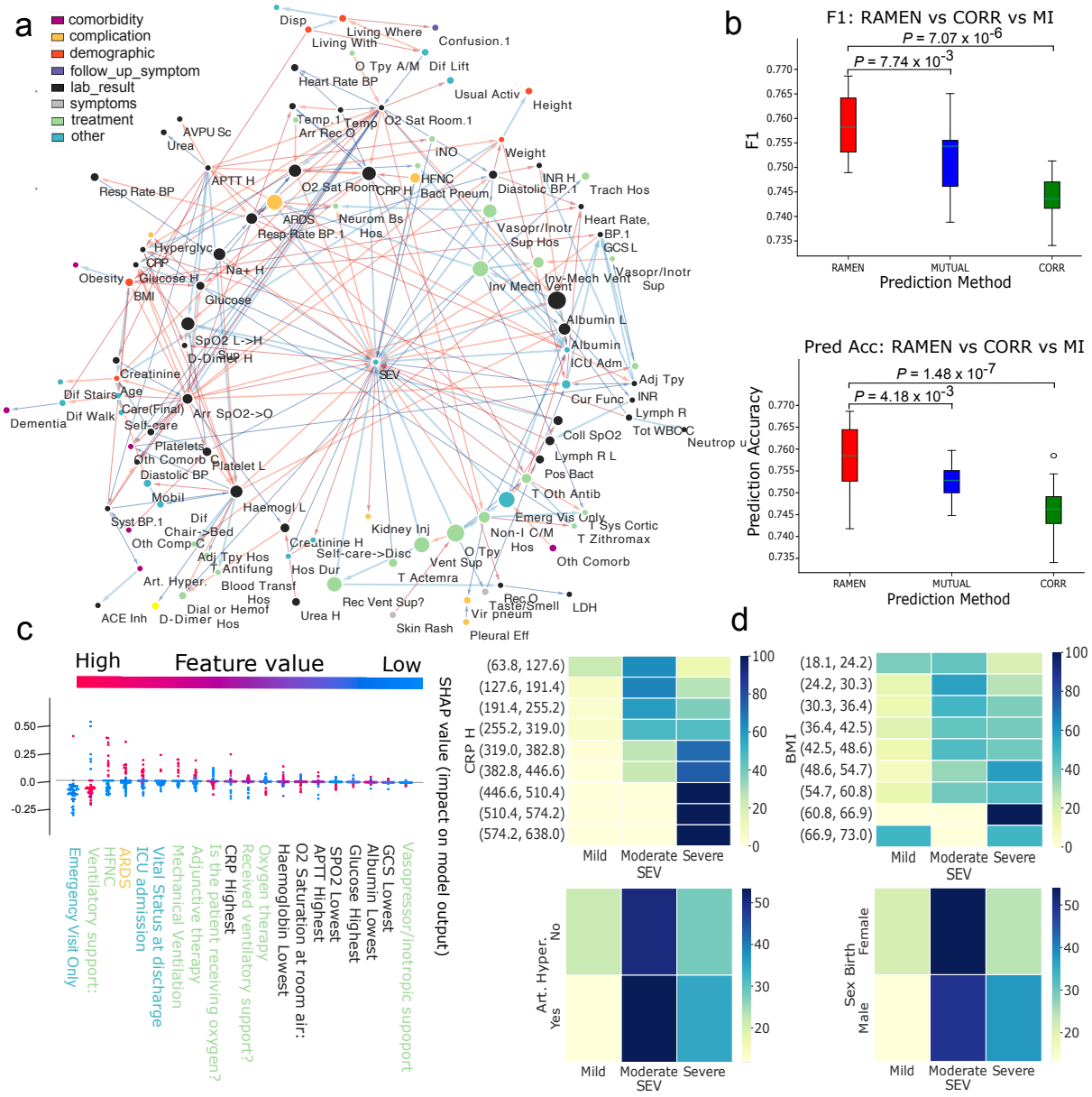

Fig. 2 RAMEN unveils indicators of COVID-19 severity in BQC19 hospitalized patient data. a, A streamlined network showcasing 231 of the most significant connections identified by RAMEN, indicative of COVID-19 severity. The full names of the variables are in Supplementary Table S1. The color and thickness of edges signify the connection strength (blue for weaker, red for stronger) based on mutual information metrics. Nodes are colored according to categories of clinical variables, with their size reflecting the strength of their correlation with COVID-19 severity. **b,** Comparison of F1 scores for predicting COVID-19 severity, contrasting RAMEN-identified indicators against those identified through mutual information (MUTUAL) and Pearson correlation (CORR) methods, with predictions made by Support Vector Machines (SVMs). A higher F1 score suggests a greater relevance of the identified variables for severity prediction. **c,** Analysis of SHapley Additive exPlanations (SHAP) values, elucidating the significance of clinical variables pinpointed by RAMEN in SVM-based predictions. These values depict the impact of variables on the model's prediction as either contributing towards a positive or negative outcome, with a consistent color scheme across the x-axis reflecting dependable predictors. **d,** Heatmaps illustrating the conditional distribution of COVID-19 severity levels (SEV) across the values of direct indicator variables, where the heatmap colors represent the proportion of patients within each severity category for given indicator values. This visual representation aids in understanding the correlation between specific clinical indicators and severity outcomes.

In our study, the RAMEN methodology was applied to the hospitalized patient cohort data from BQC19, resulting in the development of a Bayesian Network (BN). This BN delineates the complex interplay

among clinical variables and their association with COVID-19 severity (see Fig. 2a). Our analysis encompassed 2,018 hospitalized patients, with the dataset including 297 clinical variables post-cleaning. The severity of COVID-19 within this cohort was classified into three categories: "not infected or mild," "moderate," and "severe or deceased."

The clinical variable network, as inferred in relation to COVID-19 severity, aligns with findings reported in the existing literature. Among the variables identified, several key examples linked to "COVID-19 severity" include "Sex"[11, 30], "Age"[11, 30], "BMI"[31, 32], "Arterial Hypertension"[33], "ALT"[34], "C-reactive protein (CRP) (highest value)"[35], and "Albumin (lowest value)"[36]. These variables represent just a sample of the broader network, illustrating the diverse factors impacting COVID-19 severity as observed in our study.

To demonstrate the superior capability of RAMEN in identifying relevant indicators for COVID-19 outcomes compared to conventional statistical methods, we conducted benchmark analysis focused on predicting COVID-19 severity based on the early record of clinical variables. This involved training a Support Vector Machine (SVM) classifier using indicator variables from the COVID-19 severity network established by RAMEN, and two additional SVMs, each utilizing the top variables identified by mutual information and Pearson correlation methods, respectively. The effectiveness of these indicators was assessed based on the predictive performance of the SVMs. As depicted in Fig. 2b, the F1 scores and accuracy of the three SVMs, corresponding to the variables identified by each method, are presented on the x-axis. The p-values from the Wilcoxon test, comparing RAMEN with the other methods, indicate significantly higher performance for predictions based on RAMEN-identified variables. This result underscores RAMEN's ability to uncover more pertinent COVID-19 outcome indicators than traditional statistical methods. The developed classifier can also be used to predict the outcome of the disease (such as the severity of COVID-19) based on the early clinical variable records from the first month of patient care.

To assess the effectiveness of the indicators identified by RAMEN, we visualized the SHapley Additive exPlanations (SHAP) values in Fig. 2c. This visualization details how each indicator contributes to the SVM's positive or negative predictions, with the y-axis representing the contribution magnitude and the x-axis listing the RAMEN-identified indicator variables. The plot reveals a consistent pattern of value distribution (as indicated by colors) across both sides of the x-axis, with many variables situated significantly away from the axis. This indicates that these indicators are reliable predictors of outcomes, further affirming the effectiveness of the indicators identified by RAMEN.

The association between identified indicators and COVID-19 severity is further elucidated through heatmaps, as shown in Fig. 2d. These heatmaps detail the relationship between COVID-19 severity levels and the pertinent indicators identified by RAMEN. Each heatmap illustrates the variation in the percentage of patients across different severity levels in relation to the values of variables directly linked to COVID-19 severity. Generally, for variables that are strongly connected to COVID-19 severity within our reconstructed network, there is a significant shift in the distribution of severity levels corresponding to the values of these indicator variables. Conversely, for variables deemed irrelevant, the severity distribution remains largely unaffected by changes in these variables. The four heatmaps showcased underscore RAMEN's efficacy in pinpointing highly relevant indicators of COVID-19 severity.

Systematic validation of COVID-19 severity indicators identified by RAMEN using BQC19 multi-omics data

To validate the reliability of the COVID-19 severity indicators identified by RAMEN, we utilized the BQC19 multi-omics dataset. Our validation approach included a comparison of differentially expressed (DE) genes and proteins associated with each severity indicator against those associated with various levels of COVID-19 severity. This process involved examining the overlap between DE genes (from RNA sequencing data) or proteins (identified through SomaScan 5K array) related to the indicators and those distinguishing between mild and severe COVID-19 cases. The heatmaps depicted in Fig. 3a demonstrate a significant overlap of DE genes between the indicators and COVID-19 severity, revealing distinct expression patterns across the range of indicator values. These findings suggest that a common set of genes may be involved in linking these indicators to COVID-19 severity, indicating underlying biological pathways. Furthermore, the BQC19 multi-omics dataset provides insights into the biological mechanisms potentially governing these relationships. Supplementary Fig. S2 shows an additional example between "ARDS" and Severity. Pathway enrichment analysis of the "common" DE genes associated with COVID-19 severity and its primary indicators, shown in Fig. 3b, identified significant pathways including "Neutrophil degradation [37, 38],"Innate Immune System[39]," "Antimicrobial peptides[40],"

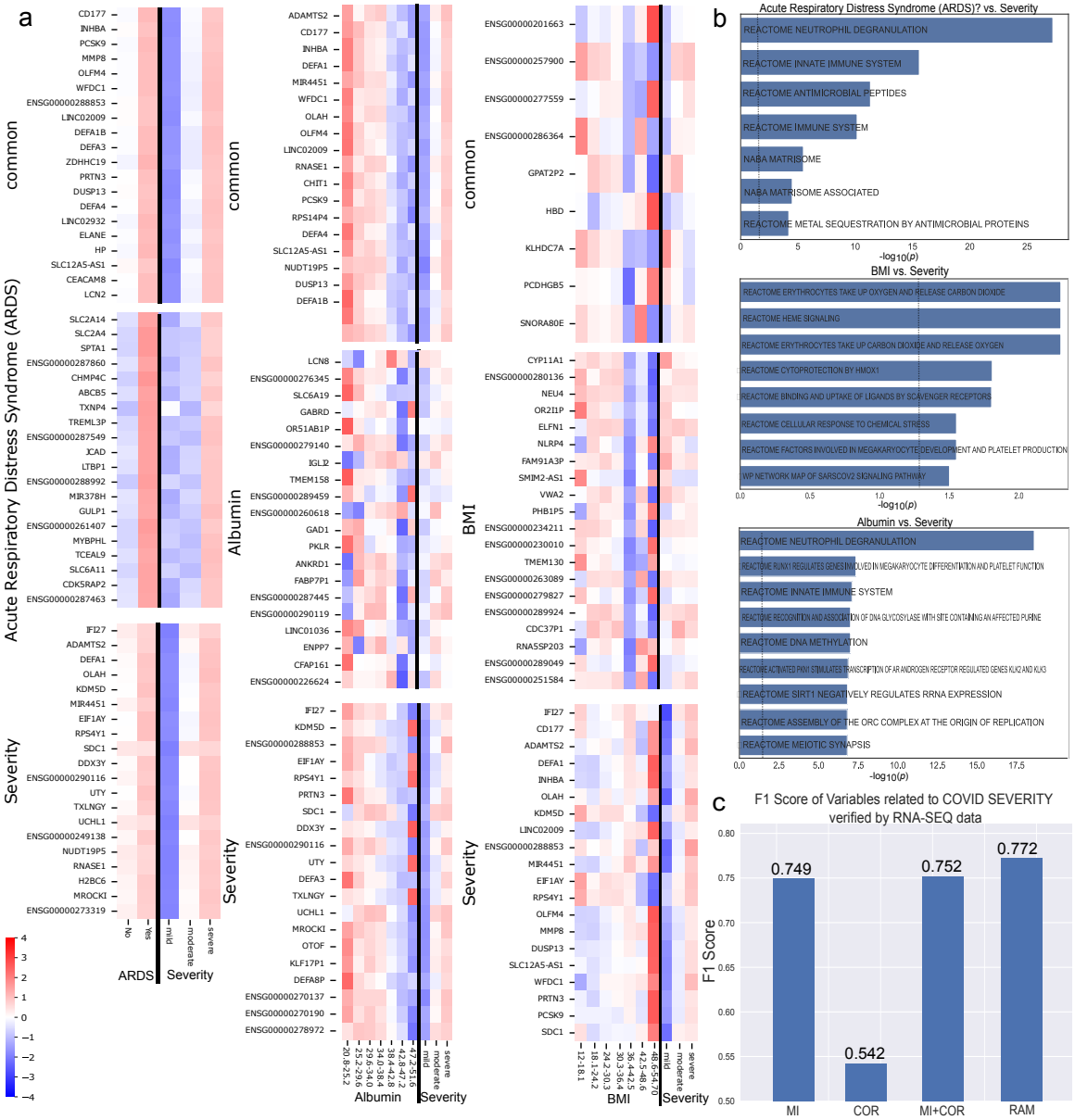

Fig. 3 Support for the COVID Severity network edges from the RNA-seq data. a, Analysis of gene expression across three groups of differentially expressed (DE) genes linked to example nodes "ARDS", "Albumin", and "BMI" that directly connected to COVID severity. For example, with "Albumin", we first pinpoint DE genes associated with Albumin variability (i.e., genes with expression changes in patients with varying Albumin levels, denoted as G_1). Next, we identify DE genes linked to COVID severity (G_2). The "Common" group in the figure represents DE genes common to both sets ($G_1 \cap G_2$); the "Albumin" group illustrates DE genes exclusive to the Albumin variable ($G_1 \cap \neg G_2$); and the "Severity" group shows DE genes unique to COVID-19 severity ($G_2 \cap \neg G_1$). **b,** Identification of the top enriched pathways for each variable based on their common DE genes with the severity variable (the "common" group). The x-axis shows the negative log 10 of FDR-corrected p-values. **c,** Validation of COVID-19 severity indicators using various methodologies. Each method on the x-axis (MI: mutual information; RAM: RAMEN; COR: Pearson correlation) classifies variables into indicators or non-indicators, with RNA-seq data providing the basis for ground truth. A variable is considered an indicator if its DE genes significantly overlap with those associated with COVID-19 severity, assessed via a hypergeometric test. The classification performance of each method is quantified using the F1 Score from verifying the variables found by each method against the ground truth.

and "Heme Signaling[41]." These pathways are implicated in modulating COVID-19 severity, suggesting mechanisms through which the indicators may influence disease severity. Severe COVID-19 is often characterized by acute respiratory distress syndrome (ARDS) associated with abnormal coagulation [42, 43]. Moreover, several studies have pointed to neutrophilia, release of their granules and neutrophil extracellular traps (NETs) as key pathological features of thrombotic complications driven by the immune system (also called immuno-thrombosis) in severe COVID-19 [3, 44–54], linking "ARDS", "Neutrophil degradation", "Innate Immune System" and "Heme Signaling". Taken together, the pathway enrichment

performed using the severity indicators identified by RAMEN is congruent with the existing literature, supporting the validity and reliability of RAMEN. We have also done this analysis using SomaScan data, which is illustrated in Supplementary Fig. S3, S4, and S5.

In addition, we performed a systematic benchmarking of RAMEN against other statistical methods, such as mutual information and Pearson correlation, to assess its effectiveness in identifying severity indicators (Fig. 3c). In this benchmarking exercise, RNA-seq data served as the basis for establishing a definitive classification of ground truth. The hypergeometric test was applied to assess the congruence between differentially expressed (DE) genes from selected indicators and those associated with COVID-19 severity, using p-values to determine statistical significance (see Methods). Clinical variables demonstrating a significant overlap of DE genes with COVID-19 severity were acknowledged as true indicators of severity for benchmarking purposes. The effectiveness of each method was assessed through the F1 score. According to Fig. 3c, correlation exhibited the lowest F1 score by a considerable margin, with Mutual Information showing significant improvement, yet still trailing behind RAMEN. The combination of correlation and mutual information was also evaluated, resulting in a marginally improved F1 score, though still not surpassing RAMEN. These results underscore RAMEN's distinctive capacity to identify severity indicators that traditional statistical methods may fail to detect.

We have also done the same benchmarking using the SomaScan data, which is shown in Supplementary Fig. S6. Similarly, correlation exhibited the lowest F1 score by a considerable margin, with mutual information showing significant improvement. The combination of correlation and mutual information resulted in an improvement from the two methods individually. However, RAMEN still demonstrated a better score than all 3 methods.

RAMEN reconstructs long COVID network from BQC19 outpatient data

We also applied RAMEN to the outpatient COVID-19 cohort to study long COVID. The outpatient dataset complements the hospitalized patients, having 1,440 patients and their 84 clinical variables. Whether or not an outpatient has long COVID is classified according to a commonly used criterion: the presence of at least one symptom at three months that persists and cannot be explained by pre-existing conditions[55]. In this cohort, based on the definition given above, long COVID frequency is 36.5% (526 out of 1440). It is important to note that this is not a random populational cohort, explaining the higher frequency. One of our objectives is to identify critical indicators for long COVID that may assist in the early diagnosis. For this study, we only used clinical variables profiled within one month after COVID-19 infection to make sure that the indicators that we found were meaningful for early diagnosis. RAMEN reconstructed a long COVID BN modeling the relationships between the clinical variables, with 36 indicators directly linked to long COVID (Fig. 4a). Many known critical variables for long COVID are captured, such as "Age"[56], "Chest"pain"[5], "Joint Pain"[57], "Runny nose"[57], and "Shortness of breath"[5]. On the other hand, irrelevant variables for long COVID, such as "chronic kidney disease", are not shown in the network. This variable, to our best knowledge, is not reported to indicate a higher risk of long COVID. As discussed in one recent study[58], an in-depth assessment of kidney outcomes in the post-acute phase of COVID-19 infection is not yet available.

While we have done the multi-omics analysis for long COVID as well, all the patients studied for long COVID are outpatients, meaning that we do not have enough multi-omics measurements for the patients to conduct a robust analysis. However, We have shown sample RNA-seq heatmaps, SomaScan heatmaps, and pathway enrichment analysis in Supplementary Fig. S7, S8, S9, S10, S11, and S12.

We demonstrate that RAMEN identifies early indicators (early records of clinical variables within the 1st month of patient care) for long COVID with enhanced precision, evidenced by the improved performance metrics (F1 score and accuracy) of an SVM classifier trained using these indicators, compared to SVM classifiers trained with indicators from other statistical methods. This indicates that RAMEN's methodology in identifying indicators specifically for long COVID effectively enhances the accuracy of predicting the disease's outcome when utilizing the SVM classifier (Fig. 4b). P-values from the one-sided Mann-Whitney U test show that RAMEN results in significantly higher accuracy and F1 score than Pearson correlation. In addition to the disease outcome prediction performance, in Fig. 4c, we also show the high feature quality of indicators found by RAMEN with SHAP values. Most indicators found are very strong early predictors of long COVID. Finally, Figure 4d illustrates the influence of specific indicators on the long COVID variable by demonstrating the shifts in the distributions of long COVID values based on the indicator values. The four variables presented are all positively associated with long COVID; for instance, the presence of "Chest Pain" (ChestP) suggests an increased risk of long COVID.

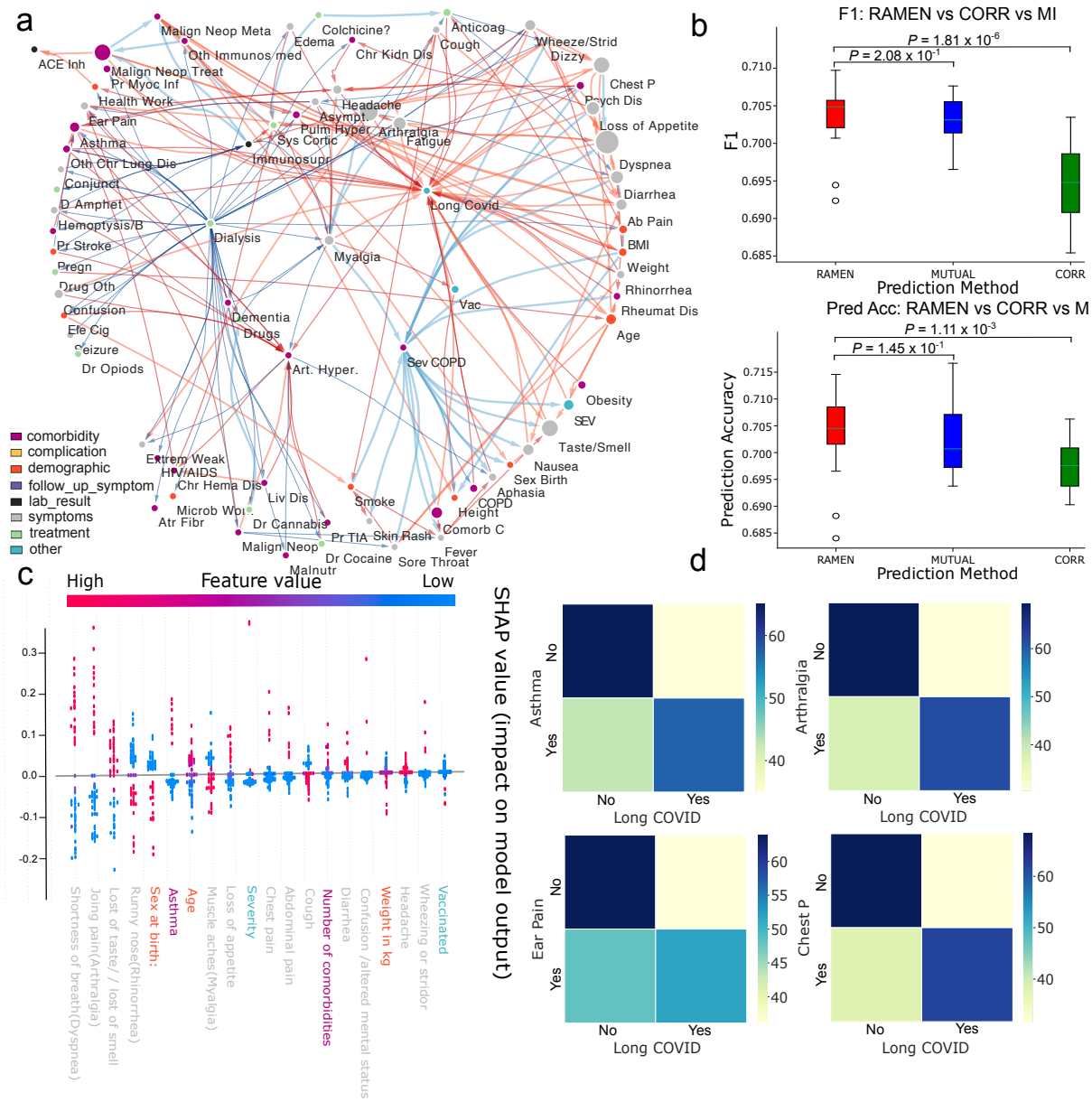

Fig. 4 RAMEN identifies long COVID indicators from BQC19 outpatient data. a, Long COVID Network constructed using RAMEN. The full names of the variables are in Supplementary Table S1. Edge colors indicate connection strength (blue for low, red for high), node colors represent types of clinical variables, and node sizes show their relevance to long COVID. **b,** F1 scores for long COVID prediction using SVMs trained on indicators identified by RAMEN, mutual information (MUTUAL), and Pearson correlation (CORR). Error bars are plotted based on 20 5-fold cross-validations and p-values are from the one-sided Mann-Whitney U test examining if one method generates significantly better results. **c,** SHAP values showing the quality of long COVID indicators identified by RAMEN. **d,** Heatmaps showing the conditional distribution of long COVID variable, conditioned on direct indicators' values.

To rigorously analyze the association between the outcome variable (long COVID) and the primary indicators, we employed Pearson's chi-square tests. The outcomes, elaborated in the Supplementary Table S2, reveal a pronounced dependency between the top indicators identified by the RAMEN network (here we showed 9) and the long COVID outcome, underscoring the robustness of these indicators in relation to the disease. To quantify the distinctiveness of these findings from potential random occurrences, we estimated the background probability that a randomly chosen variable might significantly affect the disease outcome. This involved selecting sets of 10 clinical variables from a pool excluding those identified by the RAMEN network and assessing their genomic data support. Repeating this process 100 times yielded an average of merely 1.78 out of the 10 randomly chosen variables being substantiated (equating to a background probability of 0.0178). This step was crucial to establish a benchmark for

comparison, demonstrating the specificity with which the RAMEN network's top indicators correlate with the long COVID outcome, as opposed to the general expectancy of significance from a random selection of variables. The significant p-value from the binomial test ($3.193 * 10^{-8}$) further confirms the non-randomness of the RAMEN network's findings. This stark contrast highlights the precision and relevance of the RAMEN network in pinpointing critical indicators that have a meaningful dependency relationship with the disease outcome, significantly surpassing random expectation levels.

To evaluate the robustness and consistency of RAMEN, we applied it to an independent long COVID dataset from the Lawson Health Research Institute[9], which comprises 66 patients and 27 clinical variables, to ascertain if the outcomes were consistent with our previous long COVID study using BQC19 data. The results, depicted in Supplementary Fig. S13a, illustrate the network's output using the Lawson Health Research Institute data. This dataset contains 8 variables that are common with the BQC19 dataset; of these, only 4 were identified by RAMEN as indicators of long COVID in the BQC19 dataset. These four indicators—"chest pain," "anosmia/ageusia," "dyspnea," and "headache"—were also identified by RAMEN as indicators of long COVID in the independent dataset and are highlighted in red. The identification of these indicators is supported by previous research, indicating their relevance to long COVID [5, 59–61]. Additional variables discovered in this dataset provide further insights into the pathology of long COVID. The heatmaps in Supplementary Fig. S13b demonstrate the significant impact of these indicators on the long COVID variable, with the conditional distribution of the long COVID variable showing dramatic shifts in response to these indicators.

RAMEN unveils variable relationships unreachable by mutual information or Pearson correlation for COVID-19 outcomes

To show that RAMEN can find extra information beyond simple statistical methods, especially finding network edges that cannot be reached via other methods, we tested RAMEN against Pearson correlation and mutual information. Fig. 5a and 5b show that RAMEN was able to find many edges that the other two methods were not able to find. We can see that it is especially the case in Fig. 5a, with 58.1% of the network edges being RAMEN unique. The full severity network has 79.8% being RAMEN unique. However, since Fig. 5b only shows 25% of top edges based on mutual information, there are only 22.5% of edges that are RAMEN unique.

In Fig. 5a, in the Long COVID network analysis, RAMEN identified 151 associations (edges) that were not detected by traditional correlational methods. By employing an absorbing random-walk approach integrated with a genetic algorithm, RAMEN was capable of identifying numerous associations that, while not demonstrating a strong direct correlation, play a significant role in influencing the random walk towards the absorbing node (disease outcome). This includes indicators of long COVID, such as "BMI:— Long Covid[62]". Also, it includes edges such as "COPD (emphysema, chronic bronchitis) ?— Long Covid", which is known to affect COVID severity[63]. This opens discussions to whether COPD can also be an indicator of long COVID. However, this could also be explained by the overlap of respiratory of symptoms between COPD and respiratory manifestations of long COVID, making it difficult to distinguish the etiology of the symptoms between these two conditions. Whether or not this link is a true biological association remains to be determined. Nevertheless, it demonstrates the usefulness of RAMEN in critically examining the data and the associations identified, promoting a better critical understating of pathogenesis. RAMEN can also find relationships between variables that are not directly related to long COVID, such as "Age at recruitement:— Loss of appetite ?[64]". Furthermore, indirect connections, such as "BMI:— Rheumatologic disease ?— Long Covid"[65, 66], allow us to infer the relationships between the clinical variables in the case of long COVID. While correlational methods are able to find the relation between weight and BMI significant, it is not able to suggest the rest of the sequence of relationships. In this case, we know that weight is generally a risk factor for health, hence will likely be a risk factor for long COVID. However, through RAMEN, we are able to suggest the other clinical variables that weight might provoke which might be more precise risk factors and indicators of long COVID. Accordingly, Mendelian randomization studies showed that BMI can be causally linked to rheumatoid arthritis [67, 68]. The inflammatory component of rheumatologic disease driven by BMI could interact with long-term manifestations of SARS-CoV-2 infections, generating testable hypotheses based on RAMEN-identified vectors that may help us better understand long COVID in different subgroups.

In Fig. 5b, in the compact COVID severity network, correlational methods were not able to find 52 edges that RAMEN has found (They were not able to find 739 in the complete network). To give a few,"Creatinine (HIGHEST value)— COVID Severity[69]", "Respiratory rate (associated with BP above):— COVID Severity[70]", "COVID Severity— Acute kidney injury?[71]". Acute Kidney Injury

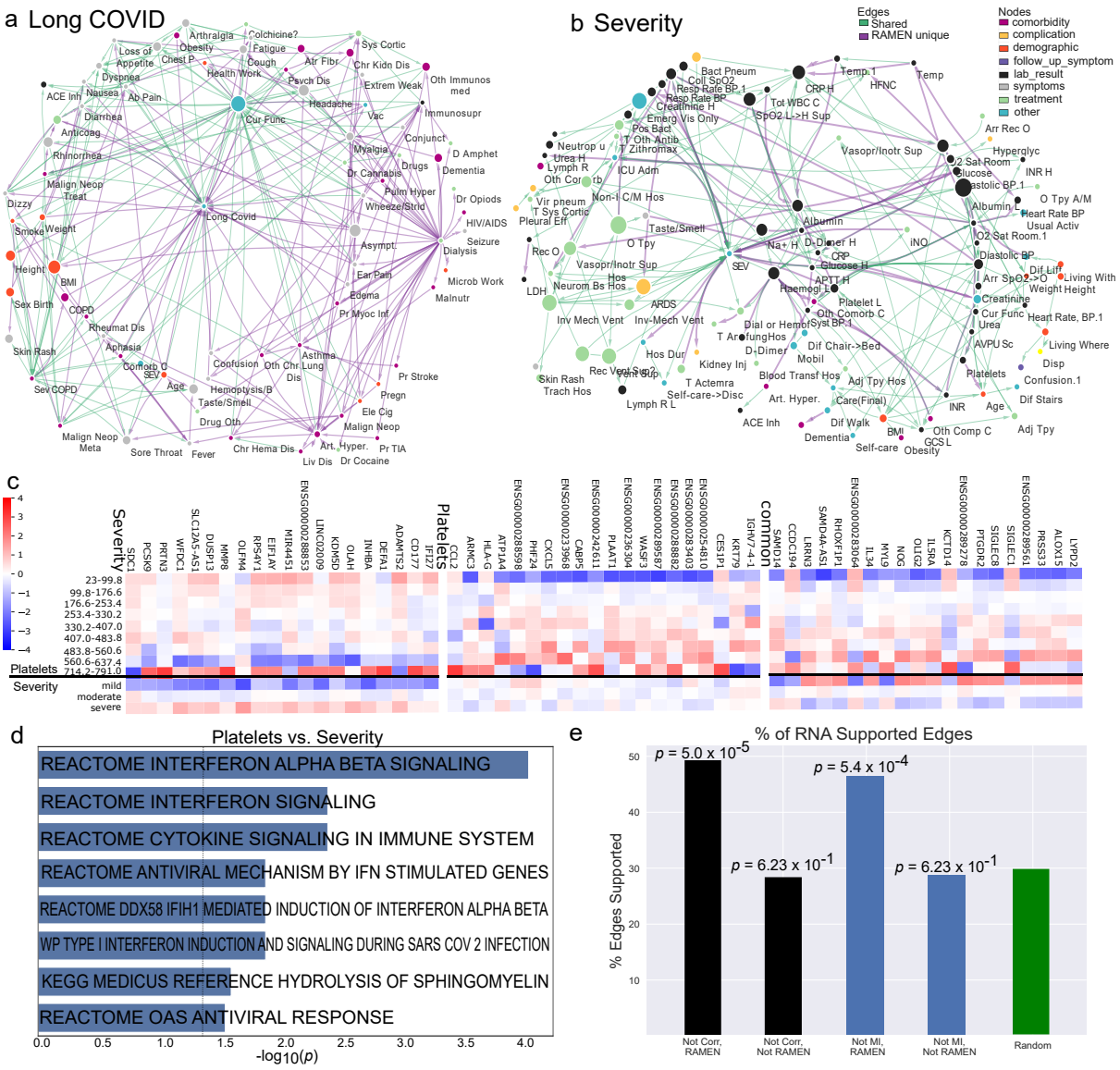

Fig. 5 RAMEN identifies effective indicator variables that cannot be found using mutual information or Pearson correlation. a, The Long COVID network where purple edges represent connections significant only to RAMEN, and green edges are also identified by mutual information or correlation. **b,** Similar network for COVID-19 severity, with purple indicating edges found exclusively by RAMEN, and green representing those also found by mutual information or correlation. The full names of the variables are in Supplementary Table S1. **c,** Heatmaps visualizing DE genes associated with "Platelets" and "COVID-19 severity". The three groups of DE genes correspond to the unique DE genes of the two variables and common DE genes. **d,** Pathway enrichment based on the common DE genes in **c**. **e,** A bar plot demonstrating RAMEN's ability to detect disease-relevant edges missed by Pearson correlation and mutual information. Using RNA-seq data as ground truth (see the Methods section* for details), among all the edges that cannot be found using Pearson correlation, the column "Not Corr, RAMEN" shows the percentage of disease-outcome-relevant edges found by RAMEN. "Not Corr, Not RAMEN" shows those that also cannot be found using RAMEN. Likewise, "Not MI, RAMEN" corresponds to the percentage of true edges missed by mutual information but found by RAMEN, and "Not MI, Not RAMEN" are the ones that are not found by both. "Random" is the performance of random selecting edges. The p-values of the binomial tests indicate that RAMEN is accurate in finding edges missed by other methods. This suggests that RAMEN has additional power in detecting disease-relevant edges compared to Pearson correlation and mutual information.

is a condition that often develops in patients affected with COVID-19, not only RAMEN was able to capture it while naive methods cannot, RAMEN was also able to capture the correct edge direction. Another edge that only RAMEN was able to capture that is worth pointing out is "Platelet (LOWEST value)— COVID Severity [72]". Platelets occupy a major function in the immune system and have been found to be an indicator of COVID severity. Naive methods not being able to find such edges show that it is lacking compared to RAMEN in identifying risk factors that can worsen COVID severity.

In Figure 5c, similar to the previous figures, we demonstrate the overlap of differentially expressed (DE) genes between "Platelets" and "COVID severity", revealing distinct expression patterns associated

with the variables' differing values. This analysis suggests the biological mechanisms underlying the predicted relationship between platelet levels and COVID-19 outcomes. In Figure 5d, a pathway enrichment analysis of the "common" DE genes related to both Platelets and COVID severity has identified several significant pathways. These pathways include "Reactome interferon alpha beta signaling" [73], "Reactome interferon signaling", "Reactome cytokine signaling in the immune system"[74], "Reactome antiviral mechanism by IFN-stimulated genes" [75], "Reactome DDX58 IFIH1-mediated induction of interferon alpha beta" [76], "WP Type I interferon induction and signaling during SARS-CoV-2 Infection"[77], "Kegg Medicus reference hydrolysis of sphingomyelin"[78], and "Reactome OAS antiviral response"[79]. Type I interferons are established immunological mediators of COVID-19 severity [80–82]. Interestingly, platelets are key regulators of coagulation, a process that can be severely disrupted in severe COVID-19 leading to life-threatening conditions[83]. These identified pathways provide a biological context for the edges predicted by the analysis, confirming their relevance to the severity of COVID-19.

We further utilized RNA-seq data to validate the edges identified by RAMEN, particularly those missed by traditional correlational methods (MI and correlation). Our validation method hinges on the principle that two variables are considered connected if there is significant overlap in their differentially expressed (DE) genes, as outlined in the Methods section. Notably, RAMEN demonstrated a significantly higher capability to identify genomics-supported edges missed by correlation methods. This contrast becomes even more pronounced when examining edges that both RAMEN and the correlational methods failed to predict; these missed edges do not exhibit any enrichment in the number of supported edges compared to random selection, indicating no significant genomics support. Fig. 5e shows that RAMEN's precision in detecting genomics-supported edges far exceeds that of random chance, as validated by the p-values from Binomial tests. Conversely, the edges missed by RAMEN showed no significant difference in support from the RNA data compared to randomly selected edges. These results affirm RAMEN's effectiveness in uncovering relevant and supported edges overlooked by conventional statistical methods.

Discussion

In this study, we address the challenge of mapping the complex network of relationships among numerous clinical variables and their effects on disease outcomes, such as COVID-19 severity and long COVID. Traditional correlation analysis methods often overlook many indirect relationships, and Bayesian network approaches, while insightful, are constrained by their high computational demand (Bayesian network structure learning is NP-hard). To overcome these limitations, we introduced RAMEN, an efficient method for Bayesian network structure learning aimed at uncovering complex interactions between clinical variables using cohort health records data. By employing random walks and a Genetic Algorithm, RAMEN effectively searches for the optimal Bayesian network structure, revealing the complex web of relationships among a wide range of variables. This method was applied to two distinct COVID-19 patient cohorts from the BQC19 dataset to delineate networks associated with COVID severity in hospitalized patients and long COVID in outpatient cohorts with repeated visits. RAMEN identified key variables linked to COVID severity and long COVID, corroborating with findings from existing research. The relationships identified by RAMEN were further validated through multi-omics data analysis within the BQC19 dataset, demonstrating the method's capability to detect complex relationships and offering insights into disease mechanisms and outcomes.

RAMEN distinguishes itself from existing methods through its ability to efficiently identify directed relationships between variables, thereby clarifying causal pathways and differentiating variables as either drivers or effectors of disease outcomes. This method enriches the analysis by discerning both direct and indirect influences on disease outcomes, shedding light on the complex web of interactions that underlie disease mechanisms. Utilizing innovative random walks and Genetic Algorithms, RAMEN overcomes the computational limitations that beset conventional Bayesian network methods, demonstrating its ability to process large datasets, such as the BQC19 dataset with 880 variables, and deliver optimized network structures swiftly. This efficiency showcases RAMEN's prowess in handling voluminous data effectively. Moreover, RAMEN introduces an approach in its network analysis by incorporating the disease outcome as a "terminal absorbing node." This strategy makes all relationships within the network conditional on reaching the disease outcome, offering a nuanced perspective that prioritizes the significance of network paths leading to the disease outcome over mere direct correlations. This strategy is particularly effective in identifying variables that significantly influence disease outcomes, providing an analytical depth not available in most existing methodologies. Additionally, RAMEN pioneers in validating inferred Bayesian networks using multi-omics data, leveraging the growing availability of such data to delve into the biological mechanisms behind diseases. This approach marks a considerable leap in the systematic

examination of reconstructed networks, effectively bridging the gap between computational analysis and biological validation. Through these advancements, RAMEN not only addresses the computational and analytical challenges posed by previous methods but also enhances the understanding of complex disease interactions.

RAMEN introduces a comprehensive toolset that fills gaps in current methodologies, enabling detailed analysis of complex relationships among clinical variables in diseases like COVID-19. This capability is pivotal for identifying biomarkers that facilitate early and accurate diagnosis, potentially leading to timely interventions and focused care by healthcare professionals to prevent complications. For instance, recognizing early joint pain as an indicator of increased risk for long COVID suggests that patients with such symptoms may benefit from prompt measures to mitigate the risk of prolonged illness. RAMEN's strength lies in its efficient and precise exploration of the interactions between clinical variables and their impact on disease outcomes. This is expected to significantly improve the strategies for early diagnosis and intervention in severe and long-term COVID cases, thereby reducing the socioeconomic burden associated with COVID-19, especially its long-term effects. While the application of RAMEN has been demonstrated primarily within the context of COVID-19 datasets in this study, the framework is designed with the flexibility to examine the network of relationships among clinical variables across various diseases.

Methods

Datasets

In this study, we utilized multi-modal data encompassing clinical variables, RNA sequencing (RNA-seq), and SomaScan 5K analysis from two distinct cohorts of COVID-19 patients (hospitalized and outpatients) within the BQC19 dataset. 380 of these patients have undergone profiling for gene expression (via RNA-seq) and protein levels (using SomaScan[84]). The clinical data collected for these COVID-19 patients encompass a range of categories, including symptoms (such as cough and muscle pain), comorbidities (like asthma and diabetes), demographic information (sex and age), laboratory test results (including white blood cell counts and ALT levels), among others. For the purpose of this analysis, we focused on hospitalized patients to examine the correlation between COVID-19 severity and various clinical indicators. This subgroup was chosen due to the relative completeness of their clinical records, particularly laboratory results, which are often missing in outpatient records. Conversely, the outpatient cohort, characterized by a higher frequency of follow-up visits, provided a valuable dataset for the investigation of long COVID.

Our analysis only included clinical variables documented within the first month of patient care, as these data points hold significant relevance for the early diagnosis of COVID-19. Consequently, the outpatient dataset incorporated a total of 1440 patients and 84 clinical variables. In addition to outpatient data, we also analyzed data from 2,018 hospitalized patients, encompassing 297 clinical variables, to construct a network model identifying markers indicative of COVID-19 severity. Further, besides the BQC19 data, we also incorporated independent COVID-19 data from the Lawson Health Research Institute, which categorizes patients into three groups: outpatients, ward patients, and ICU patients. This dataset facilitated a broader analysis across different care settings, encompassing 27 variables across 66 patients. These patients, drawn from a tertiary care system in London, Ontario, Canada, were screened and enrolled for the study, offering insights into a wide array of clinical variables categorized into symptoms, comorbidity, demographics, laboratory test results, and other relevant factors.

Bulk genomics data preprocessing

The preprocessing of the BQC19 RNA-sequencing data involved a standardized bulk RNA-seq workflow that commenced with a combined quality control and trimming step to ensure the generation of high-quality sequencing data. Initially, the raw FastQ files were subjected to FastQC (version 0.11)[85] to assess the quality of the sequencing reads, identifying potential issues such as low-quality sequences, adapter contamination, and other anomalies that could affect downstream analysis. Immediately following this quality assessment, fastp (version 0.20)[86] was employed to trim and clean the FastQ files. This step not only removed adapters and low-quality bases identified during the FastQC analysis but also ensured that the remaining reads met a high standard of quality for accurate alignment. After the quality control and trimming processes, the cleaned reads were aligned to the human reference genome hg38 using HISAT2 (version 2.2.1)[87], which efficiently produced SAM files containing detailed information about each read's alignment to the genome. These SAM files were then converted into BAM format using

SAMtools (version 1.10)[88], optimizing the data for reduced storage space and faster access, which is crucial for efficient downstream processing. Following alignment, the featureCounts (version 2.0.1) [89] software quantified the aligned reads by counting the number of reads mapping to each genomic feature, such as genes or exons, generating a comprehensive raw counts file. This file forms the basis for analyzing gene expression levels across the samples. The final step in the data preprocessing involved normalizing and transforming the raw counts data using the DESeq2(version 1.38) [90] R package. This normalization is critical for adjusting for library size variations and sequencing depth differences across samples, ensuring that the gene expression data are comparable across the entire dataset. DESeq2's normalization method prepares the data for accurate and meaningful downstream analyses, including differential expression studies, by providing a normalized set of gene expression values ready for in-depth biological interpretation. This integrated approach to RNA-seq data preprocessing—combining initial quality control with trimming, followed by alignment, quantification, and normalization—ensures that the BQC19 dataset is of the highest quality for subsequent analyses, laying the groundwork for insightful discoveries into the molecular mechanisms of COVID-19.

We also obtained SomaScan data corresponding to the circulating proteome measured by a multiplex SOMAmer affinity array (SomaLogic, 4,985 aptamers) from BQC19. SomaScan is a biotechnological protocol commercialized by the SomaLogic company (10). The SomaScan protocol comprises several levels of calibration and normalization to correct technical biases. Log2 and Z-score normalization were performed on each aptamer separately in addition to the manufacturer's provided normalized data (hybridization control normalization, intraplate median signal normalization, and median signal normalization). Since the data was analyzed by SomaLogic in two separate batches, we applied the z-score transformation separately to each batch, to reduce batch effects. These additional transformations ensure that the measured values of different aptamers are comparable and can be used in cluster analysis.

Clinical data preprocessing

In the preprocessing of the BQC19 clinical variable dataset to make it suitable for subsequent models, several steps were undertaken. Firstly, with a focus on early diagnosis, only the data from the first visit of patients with multiple visits were retained, although follow-up visits were considered for determining the presence of long COVID. For severity, for some patients, there has been more than one visit to the hospital in the first few days of showing symptoms, hence there are some measurements, such as "Temperature", that were taken during both visits. Such measurements are kept in the dataset as "Temperature.1". These variables can still be insightful, as they provide information during the early stages of symptoms of COVID-19. Secondly, variables exhibiting substantial missing values were removed. This decision was informed by creating histograms (as shown in Supplementary Fig. S14 and S15) to assess the distribution of patients having data for at least a certain number of variables, leading to the establishment of a threshold of 750 non-null values. Variables falling below this threshold for long COVID and COVID severity were excluded. Thirdly, for real-value clinical variables, discretization into bins was performed to simplify mutual information computation, with categorical variables being digitized (e.g., 0 for "no," 1 for "yes"). This approach not only facilitates mutual information analysis but also ensures a standardized preprocessing for both clinical variables and RNA-seq and SomaScan data, as described in the standardized pipeline previously outlined.

Mutual Information Calculation

To calculate mutual information effectively, it's essential to ensure that the data pairs between the two variables' vectors are both non-null. In our dataset, a common scenario is that a patient may have a recorded value for one variable but a null entry for another. Considering the independence of our patients' medical histories, employing imputation to fill these gaps is not deemed rigorous for our study. Imputation could introduce inaccuracies into our dataset. To address the missing values issue in mutual information computation, we adopted a strategy of excluding any patient data with missing entries in either of the two variables under consideration. This was accomplished through numpy array manipulation techniques applied to the two vectors. In instances where the exclusion of missing data results in an absence of patient data for the two variables (a rarity), we default to assigning a mutual information value of 0. This approach allows us to calculate mutual information between two variables accurately, notwithstanding the presence of missing data.

Build initial network structure with an absorbing random-walk-based permutation test

After the preprocessing stage, our methodology unfolds by creating a directed graph that encapsulates all clinical variables as nodes, with the connections—or edges—between these nodes weighted by the mutual information that quantifies the shared information between each pair of variables. Initially, this graph is fully connected, ensuring that each pair of nodes is linked, thus offering a comprehensive map for exploring potential relationships among all variables. To evolve this graph into a Bayesian network, which accurately reflects the true dependencies among clinical variables, we adopt a strategic approach that merges random walks with permutation testing to discern and eliminate less significant edges. This nuanced method begins with random walks across the network to evaluate the connectivity and importance of the paths between variables, with a special focus on paths leading to critical nodes such as those representing disease severity or the occurrence of long COVID.

In this refined approach, a random walk starts from a selected node X and persists until it either concludes at a designated absorbing terminal node (e.g., Severity or Long COVID) or surpasses a predetermined number of steps. A walk is categorized as "successful" if it ends at the absorbing terminal node, underscoring paths of potential significance in the disease's mechanistic understanding. The transition probability from node X to another node Y during a walk is determined by the mutual information between X and Y, normalized to ensure the sum of probabilities to one across all potential next-step nodes from X. This probabilistic framework prioritizes transitions between nodes with higher mutual information, effectively highlighting stronger relationships.

$$I(X,Y) = -\sum_x P(x)\log P(x) - \sum_y P(y)\log P(y) + \sum_x \sum_y P(x,y)\log(P(x,y)) \tag{1}$$

$$P(X|Y) = \frac{I(X,Y)}{\sum_N I(X,N)} \tag{2}$$

Concurrently, a permutation test is applied to assess the significance of these connections robustly. By randomly permuting the values among nodes and recalculating mutual information scores for these permuted networks, we generate a baseline distribution of mutual information under the hypothesis of no meaningful association between variables. Comparing the actual mutual information values against this distribution allows us to identify and subsequently trim edges that fail to statistically differentiate from what might be expected by chance. This combination of random walks for network exploration and permutation testing for evaluating edge significance presents a data-driven methodology to infer a Bayesian network. The result is a meticulously pruned network, conserving only those edges that are most indicative of genuine dependencies among clinical variables, thereby paving the way for more profound insights into the complex network of relationships that underpin disease pathology.

Bayesian network inference using Genetic Algorithm

In our study, we utilize a Genetic Algorithm (GA), a score-based structure learning method, to optimize a directed Bayesian network that captures the intricate relationships among clinical variables. This optimization builds upon an initial network framework established through random walks, which helps in identifying a substantive starting point by trimming insignificant edges. The GA is particularly suited for this task due to the NP-hard nature of Bayesian network structure learning, facilitating an efficient search for a locally optimal solution. The process begins with the creation of ten directed acyclic graph (DAG) candidates, known as parent networks. These DAGs are constructed to reflect the complex interrelations suggested by the initial network, derived from the outcomes of random walks. The GA then iteratively applies a series of genetic operations on these networks until convergence is reached. Crossover involves the random selection of pairs of networks, between which edges are exchanged. This operation promotes the exploration of new network structures by merging aspects of two parents. Mutation introduces variations within a network by randomly performing one of three possible modifications: adding one or two consecutive edges, removing one or two consecutive edges, or flipping the direction of one or two consecutive edges. This step is critical for maintaining diversity among the network population and preventing premature convergence to local optima. Scoring of each network, whether an original candidate or one newly generated through crossover and mutation, is performed using an entropy-based scoring function designed for Bayesian networks, based on mutual information, as introduced by deCampos [91]. To extend this foundational framework, we have added a regularization term specifically to constrain network complexity. This modification to the scoring function achieves a more holistic evaluation by not

only leveraging the predictive accuracy inherent in the entropy-based measure of mutual information but also incorporating a penalty for excessive complexity within the network structure. The revised scoring function is crafted to ensure that the optimization process for Bayesian network configurations effectively balances the trade-off between model accuracy and structural simplicity. The extended BN scoring function is articulated as follows:

$$\text{Score(Net)} = -\left(\sum_{X_i:\mathbb{P}_a(X_i)=\emptyset} \mathrm{H}_D(X_i) + \sum_{X_i:\mathbb{P}_a(X_i)\neq\emptyset} (\mathrm{H}_D(\{X_i\} \cup \mathbb{P}_a(X_i)) - \mathrm{H}_D(\mathbb{P}_a(X_i))) \right) - \alpha|\mathbb{E}|, \quad (3)$$

where H_D represents the entropy of a variable, reflecting its inherent uncertainty or informational content. $\mathbb{P}_a(X_i)$ denotes the set of parent variables for X_i, and $|\mathbb{E}|$ measures the network's complexity via the total number of edges, thereby penalizing excessive complexity. The hyperparameter α modulates the impact of this regularization, striking a balance between the model's simplicity and its fidelity to the data. This approach, inspired by the Kullback-Leibler (KL) divergence, strives to minimize the difference between the actual data distribution and the distribution implied by the Bayesian network model, ensuring the optimized network accurately reflects the data's underlying structure. The RAMEN pseudo-program is as Algorithm 1.

Computational validation of reconstructed network with the multi-omics data

In this section, we detail the computational approaches used to evaluate the networks formulated by RAMEN. **Validation through Clinical Variable Records:** To explore the associations between direct indicator variables and COVID-19 outcomes (severity of COVID-19 and long COVID), we constructed contingency tables that enumerate the patient counts for each pair of variable values. We applied Pearson's Chi-square test to compute a p-value, thereby assessing the dependency between these variables. **Validation via Alternative Data Modalities (Gene Expression and Protein Levels):** We began by categorizing patients based on their variable values. We then identified Differential Expression (DE) genes among these patient groups using a t-test, as facilitated by the Python package diffxpy (FDR corrected $p < 0.05$ and $|\log_2 \text{fold change}| > 0.6$)[92]. To further confirm the correlation between the variables, we employed a hypergeometric test [93] to ascertain if there is a significant overlap in DE genes. In addition, to uncover potential biological functions linking the indicator to the outcome (either severity of COVID-19 or long COVID), we examined enriched pathways associated with these clinical variables by analyzing the list of DE genes using Toppgene. This analysis aims to deepen our understanding of the biological mechanisms at play.

Support Vector Machine model to predict COVID Severity and long COVID outcomes

To evaluate the effectiveness of variables identified by RAMEN in predicting COVID outcomes, and to demonstrate the biological significance of these identified biomarkers, our approach involves a comparative analysis. Specifically, we aim to assess the predictive capability of variables selected by RAMEN against those identified through mutual information and correlation analysis. For this purpose, we utilized a Support Vector Machine (SVM) model with a linear kernel, implemented via the 'svm' module from 'sklearn'[94], to predict disease outcomes based on three sets of variables: those identified by RAMEN, the top-ranked variables according to mutual information, and the top-ranked variables according to correlation. To ensure robustness and reliability in our findings, we conducted 20 separate experiments for each set of variables, yielding 20 data points per method. Each experiment comprised a 5-fold cross-validation procedure, facilitated by the 'StratifiedKFold' class from 'sklearn.model_selection'[95], to maintain the proportion of outcomes across folds. The performance of each model was evaluated based on the average prediction accuracy and F1 score across these experiments.

To statistically ascertain whether the predictive performance of RAMEN-derived variables was significantly superior to that of variables identified by mutual information and correlation analyses, we employed a one-sided Mann-Whitney U test, available through the 'scipy.stats' module. This methodological framework not only allows us to compare the effectiveness of RAMEN in identifying predictive biomarkers but also serves to validate the biological relevance of these biomarkers by their capacity to accurately forecast disease outcomes.

Algorithm 1 Refined Bayesian Network Optimization via Genetic Algorithm

```
procedure InitializeNetwork()
    network ← Use random walks on fully connected graph
        Trim insignificant edges to form initial graph
end procedure

procedure GenerateInitialDAGs()
    DAGs ← Generate 10 initial DAGs with different seeds
end procedure

function Crossover(DAG1, DAG2)
    Exchange edges randomly between DAG1 and DAG2
end function

function Mutation(DAG)
    Randomly apply one of the following to DAG:
        1. Add one or two consecutive edges
        2. Remove one or two consecutive edges
        3. Flip the direction of one or two consecutive edges
end function

InitializeNetwork()
DAGs ← GenerateInitialDAGs()

while not Converged do
    for pair in SelectRandomPairs(DAGs) do
        Crossover(pair[0], pair[1])
    end for
    for DAG in DAGs do
        Mutation(DAG)
    end for
    Score and Select the Best DAGs for the next generation
    Converged ← CheckConvergence()
end while

procedure OutputOptimizedNetwork()
    Output the optimized Bayesian network
end procedure
```

Capturing feature importance via SHAP values

To determine the importance of variables identified by RAMEN in the SVM prediction model for COVID outcomes, we computed SHAP (SHapley Additive exPlanations) values, a technique based on game theory that assigns each feature a value indicating its contribution to the model's prediction[96]. Following the training of the SVM model with RAMEN-selected features, we utilized the SHAP library to calculate the SHAP values for these features, thereby assessing their individual impact on model predictions. This computation was facilitated by an appropriate SHAP explainer, chosen to align with the SVM's linear kernel. The resulting SHAP values were then aggregated to highlight global feature importance, providing a clear indication of how each variable influences the prediction outcome on average. By generating and analyzing SHAP summary plots, we could visually depict the rank and significance of each feature, thereby validating the predictive and biological relevance of the RAMEN-identified variables through their quantified contributions to the model's predictive accuracy.

Performance comparison via binomial test

To assess the significant differences between the proportions of supported edges identified by various methods, we employ a binomial test to calculate the p-value. Specifically, the test is formulated as $P = 1 - \text{binom.cdf}(k-1, n, p)$, where n represents the total number of edges analyzed, p is the proportion

of supported edges observed in the method with the lower rate of support, and k is the actual number of supported edges corresponding to the higher proportion, expressed as an integer. Here, binom.cdf denotes the cumulative distribution function of the binomial distribution. A p-value less than 0.05 indicates that the method with the higher proportion of supported edges is significantly more effective at predicting edges that are corroborated by genomics data, thus affirming a statistically significant difference in performance between the two methods being compared. This statistical approach allows us to rigorously evaluate and demonstrate that RAMEN's predictions are significantly more supported by genomic data compared to those derived from correlation analysis. p-values in Fig. 5 were calculated using this described approach.

Software availability

The RAMEN software, including its source code and comprehensive documentation, is freely accessible online. Users can download and explore the software from the following URL: https://github.com/mcgilldinglab/RAMEN. For the purpose of visualizing networks reconstructed by RAMEN, Cytoscape[97], a prominent tool for network visualization, was employed. Additionally, an interactive web portal has been developed to enhance the accessibility and usability of the reconstructed networks. This portal allows users to dynamically interact with the networks generated by RAMEN and can be accessed at http://dinglab.rimuhc.ca/pgm/.

Acknowledgement

This work was partially supported by CIHR PJT-180505, FRQS 295298, 295299, and NSERC RGPIN2022-04399 to JD. We thank the BQC19 initiative for granting us access to the multi-omics COVID data.

References

[1] Mofijur, M., Fattah, I.R., Alam, M.A., Islam, A.S., Ong, H.C., Rahman, S.A., Najafi, G., Ahmed, S.F., Uddin, M.A., Mahlia, T.M.I.: Impact of covid-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic. Sustainable production and consumption **26**, 343–359 (2021)

[2] Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., Agha, R.: The socio-economic implications of the coronavirus pandemic (covid-19): A review. International journal of surgery **78**, 185–193 (2020)

[3] Ding, J., Hostallero, D.E., El Khili, M.R., Fonseca, G.J., Milette, S., Noorah, N., Guay-Belzile, M., Spicer, J., Daneshtalab, N., Sirois, M., *et al.*: A network-informed analysis of sars-cov-2 and hemophagocytic lymphohistiocytosis genes' interactions points to neutrophil extracellular traps as mediators of thrombosis in covid-19. PLoS Computational Biology **17**(3), 1008810 (2021)

[4] Logue, J.K., Franko, N.M., McCulloch, D.J., McDonald, D., Magedson, A., Wolf, C.R., Chu, H.Y.: Sequelae in adults at 6 months after covid-19 infection. JAMA network open **4**(2), 210830–210830 (2021)

[5] Raveendran, A., Jayadevan, R., Sashidharan, S.: Long covid: an overview. Diabetes & Metabolic Syndrome: Clinical Research & Reviews **15**(3), 869–875 (2021)

[6] Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., *et al.*: Database resources of the national center for biotechnology information. Nucleic acids research **40**(D1), 13–25 (2012)

[7] Zuo, X., Chen, Y., Ohno-Machado, L., Xu, H.: How do we share data in covid-19 research? a systematic review of covid-19 datasets in pubmed central articles. Briefings in Bioinformatics **22**(2), 800–811 (2021)

[8] Tremblay, K., Rousseau, S., Zawati, M.H., Auld, D., Chassé, M., Coderre, D., Falcone, E.L., Gauthier, N., Grandvaux, N., Gros-Louis, F., *et al.*: The biobanque québécoise de la covid-19 (bqc19)—a

cohort to prospectively study the clinical and biological determinants of covid-19 clinical trajectories. PloS one **16**(5), 0245031 (2021)

[9] Patel, M.A., Knauer, M.J., Nicholson, M., Daley, M., Van Nynatten, L.R., Cepinskas, G., Fraser, D.D.: Organ and cell-specific biomarkers of long-covid identified with targeted proteomics and machine learning. Molecular Medicine **29**(1), 26 (2023)

[10] Vrotsou, K., Rotaeche, R., Mateo-Abad, M., Machón, M., Vergara, I.: Variables associated with covid-19 severity: an observational study of non-paediatric confirmed cases from the general population of the basque country, spain. BMJ open **11**(4) (2021)

[11] Li, X., Zhong, X., Wang, Y., Zeng, X., Luo, T., Liu, Q.: Clinical determinants of the severity of covid-19: A systematic review and meta-analysis. PloS one **16**(5), 0250602 (2021)

[12] Torres-Ruiz, J., Perez-Fragoso, A., Maravillas-Montero, J.L., Llorente, L., Mejia-Dominguez, N.R., Páez-Franco, J.C., Romero-Ramirez, S., Sosa-Hernández, V.A., Cervantes-Diaz, R., Absalon-Aguilar, A., *et al.*: Redefining covid-19 severity and prognosis: the role of clinical and immunobiotypes. Frontiers in immunology **12**, 689966 (2021)

[13] Ghayda, R.A., Lee, J., Lee, J.Y., Kim, D.K., Lee, K.H., Hong, S.H., Han, Y.J., Kim, J.S., Yang, J.W., Kronbichler, A., *et al.*: Correlations of clinical and laboratory characteristics of covid-19: a systematic review and meta-analysis. International journal of environmental research and public health **17**(14), 5026 (2020)

[14] Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

[15] Pearl, J.: Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA, pp. 15–17 (1985)

[16] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan kaufmann, 340 Pine Street, San Francisco (1988)

[17] Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge, Massachusetts (2009)

[18] Heckerman, D.E., Horvitz, E.J., Nathwani, B.N.: Toward normative expert systems: Part i the pathfinder project. Methods of information in medicine **31**(02), 90–105 (1992)

[19] Heckerman, E., Nathwani, N.: Toward normative expert systems: part ii probability-based representations for efficient knowledge acquisition and inference. Methods of Information in medicine **31**(02), 106–116 (1992)

[20] John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. arXiv preprint arXiv:1302.4964 (2013)

[21] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record **31**(1), 76–77 (2002)

[22] Cessie, S.l., Houwelingen, J.V.: Ridge estimators in logistic regression. Journal of the Royal Statistical Society Series C: Applied Statistics **41**(1), 191–201 (1992)

[23] Kohavi, R., John, G.H.: Automatic parameter selection by minimizing estimated error. In: Machine Learning Proceedings 1995, pp. 304–312. Elsevier, Tahoe City, California (1995)

[24] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences **55**(1), 119–139 (1997)

[25] Quinlan, J.R.: C4. 5: Programs for Machine Learning. Elsevier, NA (2014)

[26] Scanagatta, M., Salmerón, A., Stella, F.: A survey on bayesian network structure learning from data. Progress in Artificial Intelligence **8**, 425–439 (2019)

[27] Agrahari, R., Foroushani, A., Docking, T.R., Chang, L., Duns, G., Hudoba, M., Karsan, A., Zare, H.: Applications of bayesian network models in predicting types of hematological malignancies. Scientific reports **8**(1), 12 (2018)

[28] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. In: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, pp. 601–620 (2000)

[29] Bartoli, A., Gabrielli, F., Alicandro, T., Nascimbeni, F., Andreone, P.: Covid-19 treatment options: a difficult journey between failed attempts and experimental drugs. Internal and emergency medicine **16**, 281–308 (2021)

[30] Barek, M.A., Aziz, M.A., Islam, M.S.: Impact of age, sex, comorbidities and clinical symptoms on the severity of covid-19 cases: A meta-analysis with 55 studies and 10014 cases. Heliyon **6**(12) (2020)

[31] Caci, G., Albini, A., Malerba, M., Noonan, D.M., Pochetti, P., Polosa, R.: Covid-19 and obesity: dangerous liaisons. Journal of clinical medicine **9**(8), 2511 (2020)

[32] Chang, T.-H., Chou, C.-C., Chang, L.-Y.: Effect of obesity and body mass index on coronavirus disease 2019 severity: a systematic review and meta-analysis. Obesity Reviews **21**(11), 13089 (2020)

[33] Sisnieguez, C.E.L., Espeche, W.G., Salazar, M.R.: Arterial hypertension and the risk of severity and mortality of covid-19. European Respiratory Journal **55**(6) (2020)

[34] Kumar-M, P., Mishra, S., Jha, D.K., Shukla, J., Choudhury, A., Mohindra, R., Mandavdhare, H.S., Dutta, U., Sharma, V.: Coronavirus disease (covid-19) and the liver: a comprehensive systematic review and meta-analysis. Hepatology international **14**, 711–722 (2020)

[35] Ali, N.: Elevated level of c-reactive protein may be an early marker to predict risk for severity of covid-19. Journal of medical virology **92**(11), 2409 (2020)

[36] Paliogiannis, P., Mangoni, A.A., Cangemi, M., Fois, A.G., Carru, C., Zinellu, A.: Serum albumin concentrations are associated with disease severity and outcomes in coronavirus 19 disease (covid-19): a systematic review and meta-analysis. Clinical and Experimental Medicine **21**, 343–354 (2021)

[37] Thierry, A.R., Roch, B.: Neutrophil extracellular traps and by-products play a key role in covid-19: pathogenesis, risk factors, and therapy. Journal of clinical medicine **9**(9), 2942 (2020)

[38] Buhr, N., Parplys, A.C., Schroeder, M., Henneck, T., Schaumburg, B., Stanelle-Bertram, S., Jarczak, D., Nierhaus, A., Hiller, J., Peine, S., *et al.*: Impaired degradation of neutrophil extracellular traps: a possible severity factor of elderly male covid-19 patients. Journal of innate immunity **14**(5), 461–476 (2022)

[39] Schultze, J.L., Aschenbrenner, A.C.: Covid-19 and the human innate immune system. Cell **184**(7), 1671–1692 (2021)

[40] Elnagdy, S., AlKhazindar, M.: The potential of antimicrobial peptides as an antiviral therapy against covid-19. ACS pharmacology & translational science **3**(4), 780–782 (2020)

[41] Wagener, F.A., Pickkers, P., Peterson, S.J., Immenschuh, S., Abraham, N.G.: Targeting the heme-heme oxygenase system to prevent severe complications following covid-19 infections. Antioxidants **9**(6), 540 (2020)

[42] Asakura, H., Ogawa, H.: Covid-19-associated coagulopathy and disseminated intravascular coagulation. International journal of hematology **113**, 45–57 (2021)

[43] Helms, J., Tacquard, C., Severac, F., Leonard-Lorant, I., Ohana, M., Delabranche, X., Merdji, H., Clere-Jehl, R., Schenck, M., Fagot Gandet, F., *et al.*: High risk of thrombosis in patients with

severe sars-cov-2 infection: a multicenter prospective cohort study. Intensive care medicine **46**(6), 1089–1098 (2020)

[44] Barnes, B.J., Adrover, J.M., Baxter-Stoltzfus, A., Borczuk, A., Cools-Lartigue, J., Crawford, J.M., Daßler-Plenker, J., Guerci, P., Huynh, C., Knight, J.S., et al.: Targeting potential drivers of covid-19: Neutrophil extracellular traps. Journal of Experimental Medicine **217**(6) (2020)

[45] Bonaventura, A., Vecchié, A., Dagna, L., Martinod, K., Dixon, D.L., Van Tassell, B.W., Dentali, F., Montecucco, F., Massberg, S., Levi, M., *et al.*: Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in covid-19. Nature Reviews Immunology **21**(5), 319–329 (2021)

[46] Blasco, A., Coronado, M.-J., Hernández-Terciado, F., Martín, P., Royuela, A., Ramil, E., García, D., Goicolea, J., Del Trigo, M., Ortega, J., *et al.*: Assessment of neutrophil extracellular traps in coronary thrombus of a case series of patients with covid-19 and myocardial infarction. JAMA cardiology **6**(4), 469–474 (2021)

[47] Desilles, J.-P., Nomenjanahary, M.S., Consoli, A., Ollivier, V., Faille, D., Bourrienne, M.-C., Hamdani, M., Dupont, S., Di Meglio, L., Escalard, S., *et al.*: Impact of covid-19 on thrombus composition and response to thrombolysis: insights from a monocentric cohort population of covid-19 patients with acute ischemic stroke. Journal of Thrombosis and Haemostasis **20**(4), 919–928 (2022)

[48] Englert, H., Rangaswamy, C., Deppermann, C., Sperhake, J.-P., Krisp, C., Schreier, D., Gordon, E., Konrath, S., Haddad, M., Pula, G., et al.: Defective net clearance contributes to sustained fxii activation in covid-19-associated pulmonary thrombo-inflammation. EBioMedicine **67** (2021)

[49] Leppkes, M., Knopf, J., Naschberger, E., Lindemann, A., Singh, J., Herrmann, I., Stürzl, M., Staats, L., Mahajan, A., Schauer, C., et al.: Vascular occlusion by neutrophil extracellular traps in covid-19. EBioMedicine **58** (2020)

[50] Middleton, E.A., He, X.-Y., Denorme, F., Campbell, R.A., Ng, D., Salvatore, S.P., Mostyka, M., Baxter-Stoltzfus, A., Borczuk, A.C., Loda, M., *et al.*: Neutrophil extracellular traps contribute to immunothrombosis in covid-19 acute respiratory distress syndrome. Blood, The Journal of the American Society of Hematology **136**(10), 1169–1179 (2020)

[51] Obermayer, A., Jakob, L.-M., Haslbauer, J.D., Matter, M.S., Tzankov, A., Stoiber, W.: Neutrophil extracellular traps in fatal covid-19-associated lung injury. Disease markers **2021** (2021)

[52] Ouwendijk, W.J., Raadsen, M.P., Van Kampen, J.J., Verdijk, R.M., Von Der Thusen, J.H., Guo, L., Hoek, R.A., Van Den Akker, J.P., Endeman, H., Langerak, T., *et al.*: High levels of neutrophil extracellular traps persist in the lower respiratory tract of critically ill patients with coronavirus disease 2019. The Journal of infectious diseases **223**(9), 1512–1521 (2021)

[53] Petito, E., Falcinelli, E., Paliani, U., Cesari, E., Vaudo, G., Sebastiano, M., Cerotto, V., Guglielmini, G., Gori, F., Malvestiti, M., *et al.*: Association of neutrophil activation, more than platelet activation, with thrombotic complications in coronavirus disease 2019. The Journal of infectious diseases **223**(6), 933–944 (2021)

[54] Skendros, P., Mitsios, A., Chrysanthopoulou, A., Mastellos, D.C., Metallidis, S., Rafailidis, P., Ntinopoulou, M., Sertaridou, E., Tsironidou, V., Tsigalou, C., *et al.*: Complement and tissue factor–enriched neutrophil extracellular traps are key drivers in covid-19 immunothrombosis. The Journal of clinical investigation **130**(11), 6151–6157 (2020)

[55] Greenhalgh, T., Knight, M., Buxton, M., Husain, L., et al.: Management of post-acute covid-19 in primary care. bmj **370** (2020)

[56] Sudre, C.H., Murray, B., Varsavsky, T., Graham, M.S., Penfold, R.S., Bowyer, R.C., Pujol, J.C., Klaser, K., Antonelli, M., Canas, L.S., *et al.*: Attributes and predictors of long covid. Nature medicine **27**(4), 626–631 (2021)

[57] Aiyegbusi, O.L., Hughes, S.E., Turner, G., Rivera, S.C., McMullan, C., Chandan, J.S., Haroon, S.,

Price, G., Davies, E.H., Nirantharakumar, K., *et al.*: Symptoms, complications and management of long covid: a review. Journal of the Royal Society of Medicine **114**(9), 428–442 (2021)

[58] Bowe, B., Xie, Y., Xu, E., Al-Aly, Z.: Kidney outcomes in long covid. Journal of the American Society of Nephrology **32**(11), 2851–2862 (2021)

[59] Tana, C., Bentivegna, E., Cho, S.-J., Harriott, A.M., García-Azorín, D., Labastida-Ramirez, A., Ornello, R., Raffaelli, B., Beltrán, E.R., Ruscheweyh, R., *et al.*: Long covid headache. The Journal of Headache and Pain **23**(1), 1–12 (2022)

[60] Lippi, G., Mattiuzzi, C.: Hemoglobin value may be decreased in patients with severe coronavirus disease 2019. Hematology, transfusion and cell therapy **42**, 116–117 (2020)

[61] Huang, W., Berube, J., McNamara, M., Saksena, S., Hartman, M., Arshad, T., Bornheimer, S.J., O'Gorman, M.: Lymphocyte subset counts in covid-19 patients: a meta-analysis. Cytometry part A **97**(8), 772–776 (2020)

[62] Vimercati, L., De Maria, L., Quarato, M., Caputi, A., Gesualdo, L., Migliore, G., Cavone, D., Sponselli, S., Pipoli, A., Inchingolo, F., *et al.*: Association between long covid and overweight/obesity. Journal of Clinical Medicine **10**(18), 4143 (2021)

[63] Zhao, Q., Meng, M., Kumar, R., Wu, Y., Huang, J., Lian, N., Deng, Y., Lin, S.: The impact of copd and smoking history on the severity of covid-19: A systemic review and meta-analysis. Journal of medical virology **92**(10), 1915–1921 (2020)

[64] Donini, L.M., Savina, C., Cannella, C.: Eating habits and appetite control in the elderly: the anorexia of aging. International psychogeriatrics **15**(1), 73–87 (2003)

[65] Grainger, R., Kim, A.H., Conway, R., Yazdany, J., Robinson, P.C.: Covid-19 in people with rheumatic diseases: risks, outcomes, treatment considerations. Nature Reviews Rheumatology **18**(4), 191–204 (2022)

[66] Iannone, F., Lopalco, G., Rigante, D., Orlando, I., Cantarini, L., Lapadula, G.: Impact of obesity on the clinical outcome of rheumatologic patients in biotherapy. Autoimmunity reviews **15**(5), 447–450 (2016)

[67] Bae, S.-C., Lee, Y.H.: Causal association between body mass index and risk of rheumatoid arthritis: a mendelian randomization study. European journal of clinical investigation **49**(4), 13076 (2019)

[68] Zhao, S.S., Holmes, M.V., Zheng, J., Sanderson, E., Carter, A.R.: The impact of education inequality on rheumatoid arthritis risk is mediated by smoking and body mass index: Mendelian randomization study. Rheumatology **61**(5), 2167–2175 (2022)

[69] Ok, F., Erdogan, O., Durmus, E., Carkci, S., Canik, A.: Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of covid-19 patients. Journal of medical virology **93**(2), 786–793 (2021)

[70] Mudatsir, M., Fajar, J.K., Wulandari, L., Soegiarto, G., Ilmawan, M., Purnamasari, Y., Mahdi, B.A., Jayanto, G.D., Suhendra, S., Setianingsih, Y.A., et al.: Predictors of covid-19 severity: a systematic review and meta-analysis. F1000Research **9** (2020)

[71] Hirsch, J.S., Ng, J.H., Ross, D.W., Sharma, P., Shah, H.H., Barnett, R.L., Hazzan, A.D., Fishbane, S., Jhaveri, K.D., Abate, M., *et al.*: Acute kidney injury in patients hospitalized with covid-19. Kidney international **98**(1), 209–218 (2020)

[72] Barrett, T.J., Bilaloglu, S., Cornwell, M., Burgess, H.M., Virginio, V.W., Drenkova, K., Ibrahim, H., Yuriditsky, E., Aphinyanaphongs, Y., Lifshitz, M., *et al.*: Platelets contribute to disease severity in covid-19. Journal of Thrombosis and Haemostasis **19**(12), 3139–3153 (2021)

[73] Gill, S.E., Dos Santos, C.C., O'Gorman, D.B., Carter, D.E., Patterson, E.K., Slessarev, M., Martin,

C., Daley, M., Miller, M.R., Cepinskas, G., *et al.*: Transcriptional profiling of leukocytes in critically ill covid19 patients: implications for interferon response and coagulation. Intensive care medicine experimental **8**, 1–16 (2020)

[74] Rabaan, A.A., Al-Ahmed, S.H., Muhammad, J., Khan, A., Sule, A.A., Tirupathi, R., Mutair, A.A., Alhumaid, S., Al-Omari, A., Dhawan, M., *et al.*: Role of inflammatory cytokines in covid-19 patients: A review on molecular mechanisms, immune functions, immunopathology and immunomodulatory drugs to counter cytokine storm. Vaccines **9**(5), 436 (2021)

[75] Loganathan, T., Ramachandran, S., Shankaran, P., Nagarajan, D., *et al.*: Host transcriptome-guided drug repurposing for covid-19 treatment: a meta-analysis based approach. PeerJ **8**, 9357 (2020)

[76] Potamias, G., Gkoublia, P., Kanterakis, A.: The two-stage molecular scenery of sars-cov-2 infection with implications to disease severity: An in-silico quest. Frontiers in Immunology **14** (2023)

[77] Park, A., Iwasaki, A.: Type i and type iii interferons–induction, signaling, evasion, and application to combat covid-19. Cell host & microbe **27**(6), 870–878 (2020)

[78] Kornhuber, J., Hoertel, N., Gulbins, E.: The acid sphingomyelinase/ceramide system in covid-19. Molecular Psychiatry **27**(1), 307–314 (2022)

[79] Wauters, E., Van Mol, P., Garg, A.D., Jansen, S., Van Herck, Y., Vanderbeke, L., Bassez, A., Boeckx, B., Malengier-Devlies, B., Timmerman, A., *et al.*: Discriminating mild from critical covid-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. Cell research **31**(3), 272–290 (2021)

[80] Zhang, Q., Bastard, P., Cobat, A., Casanova, J.-L.: Human genetic and immunological determinants of critical covid-19 pneumonia. Nature **603**(7902), 587–598 (2022)

[81] Matuozzo, D., Talouarn, E., Marchal, A., Zhang, P., Manry, J., Seeleuthner, Y., Zhang, Y., Bolze, A., Chaldebas, M., Milisavljevic, B., *et al.*: Rare predicted loss-of-function variants of type i ifn immunity genes are associated with life-threatening covid-19. Genome Medicine **15**(1), 22 (2023)

[82] Su, C., Rousseau, S., Emad, A.: Identification of transcriptional regulatory network associated with response of host epithelial cells to sars-cov-2. Scientific Reports **11**(1), 23928 (2021)

[83] Sciaudone, A., Corkrey, H., Humphries, F., Koupenova, M.: Platelets and sars-cov-2 during covid-19: Immunity, thrombosis, and beyond. Circulation Research **132**(10), 1272–1289 (2023)

[84] Gold, L., Walker, J.J., Wilcox, S.K., Williams, S.: Advances in human proteomics at high scale with the somascan proteomics platform. New biotechnology **29**(5), 543–549 (2012)

[85] Andrews, S., et al.: FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom (2010)

[86] Chen, S., Zhou, Y., Chen, Y., Gu, J.: fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics **34**(17), 884–890 (2018)

[87] Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L.: Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nature biotechnology **37**(8), 907–915 (2019)

[88] Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., *et al.*: Twelve years of samtools and bcftools. Gigascience **10**(2), 008 (2021)

[89] Liao, Y., Smyth, G.K., Shi, W.: featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**(7), 923–930 (2014)

[90] Love, M., Anders, S., Huber, W.: Differential analysis of count data–the deseq2 package. Genome Biol **15**(550), 10–1186 (2014)

[91] De Campos, L.M., Friedman, N.: A scoring function for learning bayesian networks based on mutual information and conditional independence tests. Journal of Machine Learning Research **7**(10) (2006)

[92] Kang, H.C., Kim, I.-J., Park, J.-H., Shin, Y., Ku, J.-L., Jung, M.S., Yoo, B.C., Kim, H.K., Park, J.-G.: Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. Clinical Cancer Research **10**(1), 272–284 (2004)

[93] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[94] Kramer, O., Kramer, O.: Scikit-learn. Machine learning for evolution strategies, 45–53 (2016)

[95] Prusty, S., Patnaik, S., Dash, S.K.: Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. Frontiers in Nanotechnology **4**, 972421 (2022)

[96] Marcílio, W.E., Eler, D.M.: From explanations to feature selection: assessing shap values as feature selection mechanism. In: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 340–347 (2020). https://doi.org/10.1109/SIBGRAPI51738.2020.00053

[97] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**(11), 2498–2504 (2003)