Identifying a proteomic signature for COVID-19 outcomes

Chen-Yang Su, School of Computer Science
McGill University, Montreal, Quebec, Canada

April, 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Computer Science

©Chen-Yang Su, 2022

Abstract

The reason behind why some people have severe COVID-19 and others develop asymptomatic infection is not well understood. Understanding this phenomenon may help to provide insights into the causes of severe COVID-19 and identify individuals at high risk of suffering from it. Currently, the biological mechanisms involved in severe COVID-19 are not fully understood, and it remains difficult to predict the clinical outcomes of infected patients. The mechanisms underlying COVID-19 infection can be studied through host genetics and proteomics methods and could enable the development of better therapeutics to treat and reduce the complications of COVID-19. One way to quickly assess thousands of potential biological mechanisms leading to severe COVID-19 is to use circulating proteins. These proteins are privileged drug targets because they are found in circulation and can be easily measured with a blood test. To approach this problem, we measured the abundance of 4,701 circulating human proteins in two independent cohorts totaling 986 individuals. We then used these measures along with several clinical factors relevant to COVID-19 to develop predictive models for severe COVID-19 in 417 participants and validated these models in a separate cohort of 569 people. We also tested whether the model including circulating proteins better predicted severe COVID-19 compared to models including clinical risk factors without proteins. A baseline model including age and sex provided an area under the receiver operator curve (AUC) of 65% in the test cohort. The addition of 92 proteins from the 4,701 unique protein abundances improved the AUC to 88% in the training cohort, and to 86% in the test cohort, suggesting good generalizability. All 92 selected proteins were enriched for cytokine receptors, but more than half of the enriched pathways were not immune-related. In summary, these results suggest that measuring circulating proteins can allow reasonably accurate predictions of clinical outcomes of COVID-19. In order to understand how to incorporate protein measurement and translate findings into clinical care, further research will be needed.

Abrégé

La raison pour laquelle certaines personnes développent une COVID-19 sévère et d'autres une infection asymptomatique demeure incomprise. Une meilleure compréhension de ce phénomène pourrait aider à mieux cerner les causes de la COVID-19 sévère et identifier les individus à haut risque d'en souffrir. Actuellement, les mécanismes biologiques impliqués dans la COVID-19 sévère ne sont pas entièrement comprises, et il demeure difficile de prédire les issus cliniques des patients infectés. Les mécanismes sous-tendant l'infection au COVID-19 peuvent étudiés grâce à des méthodes de génétique et de protéomique de l'hôte, et pourrait permettre le développement de meilleurs traitements pour traiter et réduire les complications de la COVID-19. Une façon d'évaluer rapidement des milliers de mécanismes biologiques potentiels conduisant à un COVID-19 grave consiste à utiliser des protéines circulantes. Ces protéines sont des cibles pharmaceutiques à potentiel élevés car elles se trouvent dans la circulation, et facilement mesurables avec une analyse sanguine. Ainsi, l'études es protéines circulantes pourrait permettre l'élaboration de modèles prédictifs de la COVID-19 sévère et faciliter le développement de meilleures thérapies. Pour résoudre ce problème, nous avons mesuré l'abondance de 4 701 protéines humaines en circulation dans deux cohortes indépendantes totalisant 986 individus. Nous avons ensuite utilisé ces mesures ainsi que plusieurs facteurs cliniques pertinents à la COVID-19, et avons développé des modèles de prédiction pour la COVID-19 sévère chez 417 participants. Nous avons ensuite validé ces modèles dans une cohorte distincte de 569 personnes. Nous avons aussi testé si le modèle incluant des protéines circulantes prédisait mieux la COVID-19 sévère par rapport aux modèles incluant les facteurs de risque clinique sans protéines. Un modèle de référence incluant l'âge et le sexe du patient obtient une aire sous la courbe récepteur-opérateur (AUC) de 65% dans la cohorte test. L'ajout de 92 protéines parmi les 4 701 mesurées augmente l'AUC à 88% dans la cohorte d'entraînement, et à 86% dans la cohorte test, suggérant une bonne généralisabilité. L'ensemble des 92 protéines sélectionnées est enrichi en récepteurs de cytokines, mais plus de la moitié des voies enrichies n'étaient pas non plus directement liées au système immunitaire. En résumé, ces résultats suggèrent que la mesure des protéines circulantes peut permettre des prédictions raisonnablement précises des issus cliniques de la COVID-19. Afin de comprendre comment intégrer la mesure des protéines et traduire les résultats en soins cliniques, des recherches supplémentaires seront nécessaires.

Contributions

The main contribution of this thesis is the use of machine learning in the evaluation of one of the largest datasets of circulating proteins in the world in COVID-19 patients to predict whether an individual will be more susceptible to COVID-19 infection or develop severe disease. In summary, the two main contributions of this work are as follows: 1) to utilize a dataset composed of the largest set of circulating protein biomarkers measured in COVID-19 patients collected over the course of disease progression to design an accurate, clinically relevant multiprotein machine learning model for the prediction of COVID-19 adverse outcomes. 2) To externally validate the robustness of the model in a separate dataset from collaborators at Mount Sinai in New York, who have generated the same proteomic profiling and clinical outcomes.

Preliminary parts of the work in this thesis has been accepted as an abstract at the American Society of Human Genetics (ASHG) in Oct 2021 and presented as a poster presentation by the author of this thesis who is a first co-author of the work along with two other first authors: Drs. Sirui Zhou, and Edgar Gonzalez-Kozlova, and a senior author: Dr. Brent Richards. Major parts of this work has also been submitted to Scientific Reports and is under review at the current preparation stage of this thesis. Both Drs. Sirui Zhou, Edgar Gonzalez-Kozlova, and Brent Richards have acknowledged the use of the abstract and the manuscript in this thesis submission.

Acknowledgements

First and foremost, I would like to thank everyone that has been a part of my life over the past two years of my Master's (Sept 2020 – June 2022). Due to COVID-19, my graduate school experience was not the same as I had hoped and dreamed of, but it would not have been as stimulating, fun, and memorable without those that helped me and were with me along the way. Thank you to my supervisors Dr. Joelle Pineau and Dr. Brent Richards for their advice and supervision. Dr. Joelle Pineau provided helpful discussions and guidance on the use of machine learning properly. Dr. Brent Richards provided crucial expertise and knowledge in the field of human genetics and led me to the important problems that should be tackled to improve clinical care.

I am deeply grateful for Dr. Sirui Zhou who was a postdoc in Dr. Brent Richards' lab until Dec 2021 (now assistant professor in human genetics at McGill University) for her support and guidance throughout my entire Master's. She was instrumental to my success and taught me so much in the field of genetics and proteomics. I would not be where I am right now without her. I would like to thank Marc-André Legault who was extremely helpful in the preparation of this thesis and acted as an amazing mentor. His expertise in both genomics and machine learning was greatly appreciated.

I would like to thank members of the Richards' Lab that have guided and supported me in my journey into human genetics and for their helpful discussions along the way: Guillaume Butler-Laporte, Tomoko Nakanishi, Yiheng Chen, Tianyuan Lu, Kevin Liang, Vince Forgetta, Yossi Farjoun, David Morrison, Laetitia Laurent, and many others.

Thank you to the the Biobanque Québécoise de la COVID-19, SomaLogic (Clare Paterson, Mike Hinterberg), and all those at the Icahn School of Medicine at Mount Sinai in New York City for a great collaboration, of which this thesis would not have been possible without.

I am deeply appreciative of Dr. Wonseok Jeon who was a postdoc in Dr. Joelle Pineau's lab until April 2021 (now a Staff Machine Learning Engineer at Qualcomm AI Research) and supervised me in the first year of my Master's. He is not only a mentor but a close friend who has taught me so much in terms of research and what graduate school is all about.

I would also like to thank my family, especially both my parents, for pushing me through the difficult times and supporting me along the way. Thank you to my friends, from the west coast to east coast of Canada, for your time spent with me outside of the lab.

Finally, I want to thank the Lady Davis Institute, Mila - Quebec Artificial Intelligence Institute, and Compute Canada for providing me with an exceptionally high-quality environment and compute power for conducting research. Thank you to the Lady Davis Institute and TD Bank for supporting me with an LDI/TD Bank Studentship.

Table of Contents

Abstract i

Abrégé iii

Contributions v

Acknowledgements vi

List of Figures xii

List of Tables xiii

List of terms **xiv**

1 Introduction **1**

2 Background and Literature Review **5**

2.1 Basic Epidemiological Terminology 8

2.2 Human Genetics 9

2.3 Genome-Wide Association Studies 10

2.4 Multiple Testing 12

2.4.1 Bonferroni Correction: Controlling the Familywise Error Rate 13

2.4.2 Benjamini-Hochberg Procedure: Controlling the False Discovery Rate (FDR) 14

3 Primer on Machine Learning **15**

3.1 Regression Models 16

3.1.1 Linear Regression 17

3.1.2 Logistic Regression 19

3.2 Bias-Variance Trade-off 22

3.3 Sparse Models and Regularization 23

3.3.1 L1 Regularization 23

3.3.2 L2 Regularization 24

3.3.3 Elastic Net Regularization 25

3.4 Cross-Validation 26

3.5 Model Evaluation and Assessment 27

3.5.1 Confusion Matrix 27

3.5.2 Precision/Recall Curves; Sensitivity/Specificity 29

3.5.3 Area under the Receiver Operating Characteristic 29

3.5.4 Youden's J Statistic 30

3.6 Dimensionality Reduction 31

4 Circulating Proteins to Predict Adverse COVID-19 Outcomes 33

4.1 Methods 34

4.1.1 Cohorts 34

4.1.2 Demographic, Sample Processing, and Clinical Variable Definitions . 35

4.1.3 Proteomic Measurement using the Somascan Platform 37

4.1.4 SomaLogic Quality Control and Normalization 38

4.1.5 Data Preprocessing 39

4.1.6 Curation of Samples from the Longitudinal Dataset 39

4.1.7 COVID-19 Outcome Definitions 40

4.1.8 Multivariable Logistic Regression 40

4.1.9 Regularized Logistic Regression Models 41

4.1.10 Using Elastic Net to Select Uncorrelated Proteins 42

4.1.11 Cross-Validation and Hyperparameter Tuning 43

4.1.12 Model Testing 45

4.1.13 Protein Correlations 45

4.1.14 Pathway Enrichment Analyses 46
4.1.15 Sensitivity Analyses 46
4.2 Experimental Results 47
4.2.1 Cohorts 47
4.2.2 Association of Protein Abundance with COVID-19 Outcomes 49
4.2.3 Model Selection and Performance Using LASSO 50
4.2.4 Training and Performance of Models using Elastic Net 54
4.2.5 Proteins Prioritized by LASSO to Predict COVID-19 Severity Outcomes 57
4.2.6 Using Clinical Risk Factors to Predict COVID-19 Outcomes 59
4.3 Discussion 60
4.3.1 Prediction of COVID-19 Outcomes 61
4.3.2 Prediction of COVID-19 Outcomes with Circulating Proteins 62
4.3.3 Inclusion of Secondary Cohort 62
4.3.4 Circulating Cytokines and Chemokines 63
4.3.5 Limitations 63
4.4 Miscellaneous 64
4.4.1 Ethics declarations 64
4.4.2 Data Availability 64
4.4.3 Author Contribution 65
4.4.4 Acknowledgements 65

5 Conclusion and Future Work **79**

5.1 Conclusion 79
5.2 Future Work 80

A Supplementary Tables **95**

List of Figures

2.1 OMICS technologies: moving from DNA toward phenotype 10

2.2 Manhattan plot 12

3.1 Plot of logistic function 22

3.2 Illustration of Cross-Validation 27

3.3 Confusion Matrix Example 28

3.4 Area Under the Receiving Operating Curve 30

4.1 Overall study design 49

4.2 Violin plots showing the age and sex distribution of the BQC19 training cohort for severe and critical COVID-19. 50

4.3 Volcano plots of severe and critical COVID-19. 51

4.4 AUC score as a function of regularization strength during L_1 regularized logistic regression training 52

4.5 L_1 regularized logistic regression training and testing AUC results for severe COVID-19 and critical COVID-19 53

4.6 Two-by-two contingency table results for predicting severe and critical COVID-19 using the protein model and baseline model (red) 55

4.7 Elastic net logistic regression hyperparameter search results 56

4.8 Spearman's rank correlations between the 92 proteins within the protein model trained to predict the severe COVID-19 outcome. 66

4.9 Feature importance and correlation of SOMAmers selected in the protein model to predict severe and critical COVID-19. 67
4.10 Spearman's rank correlations between the 67 proteins within the protein model trained to predict the critical COVID-19 outcome. 69
4.11 Spearman's rank correlations between the nonoverlapping proteins predicting severe and critical COVID-19 outcomes. 70
4.12 Unsupervised clustering by uniform manifold approximation and projection (UMAP) on 4,984 SOMAmer reagents in the BQC19 cohort. 71
4.13 Pathway analysis on common proteins best predicting severe and critical COVID-19 outcomes. 72
4.14 Main Analysis plus 6 clinical risk factors AUC scores 73
4.15 Main Analysis plus 6 clinical risk factors: model coefficients. 74
4.16 Main Analysis plus 7 clinical risk factors AUC scores 76
4.17 Main Analysis plus 7 clinical risk factors: model coefficients. 77

List of Tables

4.1 Demographic characteristics of the participating cohorts. 36
4.2 Notes for Table 4.1 37
4.3 Rate of missingness for seven clinical variables in the BQC19 training cohort. 48
4.4 Training AUC comparison. 61

Terminology

complex diseases diseases that are influenced by more than one gene along with factors such as the environment and lifestyle. 10

genome-wide association studies study design used to determine the association between genetic variants and a trait or disease. 10

genotype microarrays platform to identify variants in multiple regions in the genome that are associated with a disease. 11

heritability the degree of variation that is due to genetic variation. 10

indels any insertion or deletion of nucleotides in the genome. 11

information leakage when a model is trained on information that should not be available in the training set. 15

linkage disequilibrium the non-random association of alleles at two or more loci. 12

linkage disequilibrium pruning removing loci that are in high pairwise linkage disequilibrium. 12

mendelian disorders Diseases caused by mutations or alterations in a single gene. 10

prediction bias the difference between the average model prediction and the average of the true targets (actual values). 15

single nucleotide polymorphisms Variation in the genome due to a single nucleotide change. 11

single nucleotide variants a single nucleotide change in the DNA sequence. 10, 11

Chapter 1

Introduction

The COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) remains to be one of the most difficult challenges countries around the world face today [1]. In particular, the emergence of novel strains may reduce the efficacy of vaccines and lead to more breakthrough cases. Currently, few disease specific therapies exist for treating severe SARS-CoV-2 infection [2] and the prediction of which individuals will experience these adverse outcomes is difficult.

Proteins are complex biomolecules which play major roles in the human body such as modulating disease, regulating key biochemical processes, and maintaining repair functionality. Within the human genome, nearly 30,000 proteins are encoded by 19,000 genes [3]. The expression patterns and regulation of proteomic biomarkers may enable novel discoveries into how immune responses to viral infection occur [4]. Determining proteins that play key roles during viral infection can provide useful information to identify groups of individuals who are at risk for severe COVID-19 outcomes, such as intubation and death. We hypothesize that large-scale blood proteomic data in COVID-19 patients can be used to predict the course and outcome of the disease. Further, changes in protein abundances in patients during infection may be driven by the host genome and the identification of these drivers may assist drug development. This is because human genetic findings can help to pinpoint critical control points in disease etiology.

A remarkable feature of COVID-19 disease is its highly variable clinical course, where some individuals manifest severe disease or death, and others remain asymptomatic. Several clinical and genetic risk factors explain a proportion of these outcomes [4–8], yet most of the host biological causes of these adverse COVID-19 outcomes remain unknown.

Recent reports have identified some of the biologic pathways influencing risk of adverse COVID-19, such as immune responses [9–12], interferon pathways [13–15], and T-cell dysfunction [16, 17]. However, many such studies have focused on narrow sets of pre-selected cytokines. To expand this analysis, proteomic studies can rapidly assess thousands of potential biomarkers associated with the severity of COVID-19 through the measurement of blood circulating proteins. Such circulating proteins may be useful because they can help to identify pathways influencing severity of disease. They may also identify individuals at high risk of a severe COVID-19 clinical course. Similarly, circulating proteomic biomarkers have recently been shown to serve as predictors of other common diseases [18–24] including cardiovascular disease. They are also relevant in drug discovery because they are generally more accessible to pharmacological manipulation than intracellular proteins [25–29]. Thus, understanding the circulating proteins associated with adverse COVID-19 outcomes may be helpful to address major challenges raised by the current pandemic [16,30–41].

There are two main objectives of this research work: 1) to utilize a dataset composed of the largest group of circulating protein biomarkers for COVID-19 patients collected over the course of disease progression to design an accurate, clinically relevant multiprotein machine learning model for the prediction of COVID-19 adverse outcomes. 2) To externally validate the robustness of the model in a separate dataset from collaborators at Mount Sinai in New York, who have generated the same proteomic profiling and clinical outcomes. These two objectives are described and presented in further detail in chapter 4 of this thesis.

For this project, we used already measured and publicly available circulating protein levels of patients enrolled in the Biobanque Québécoise de la COVID-19 (BQC19) [42].

Protein levels were measured using the SomaLogic SomaScan platform [43], and to date, this data consists of 1,957 blood samples collected from 1,253 individuals across two batches with 4,701 protein measurements for each sample. Due to the earlier completion date of the work presented by the first two objectives of this thesis (chapter 4), only half of the samples (1,039 samples) from 504 individuals were available.

In chapter 4 of this thesis, the overarching goal is to identify a proteomic signature for COVID-19 outcomes. In the first objective of this thesis, COVID-19 outcomes are classified into two groups based on different levels of disease severity, and 4,984 nucleic acid aptamers (SOMAmer reagents) [44] targeting 4,701 proteins measured on the SomaLogic platform are leveraged along with clinical and demographic variables to predict these outcomes using dimensionality reduction models. Previous studies have shown the effectiveness of leveraging a similar set of proteomic biomarkers as prognostic tools for developing predictive models related to COVID-19 infection [19] and other disease statuses [18,20]. Here, we developed and utilized machine learning algorithms and models such as Least Absolute Shrinkage and Selection Operator (LASSO), Ridge regression, and elastic-net logistic regression to predict COVID-19 outcomes. One challenge to overcome in order to produce accurate predictions is to reduce the number of proteins to a manageable amount before inputting them into the models. Methods such as Principal Component Analysis (PCA), LASSO, and uniform manifold approximation and projection (UMAP) were used for this task of dimensionality reduction. In addition, we used False Discovery Rate (FDR) corrected P values from logistic regression to filter out non-significant proteins.

In the second objective of this thesis, we tested the external validity of our findings by replicating our results in the Icahn School of Medicine at Mount Sinai COVID-19 biobank [45], which consists of 4,701 proteins measured from 569 individuals. To do so, we selected the optimal model trained in the BQC19 cohort and applied it to the Mount Sinai COVID-19 biobank. We then tested the sensitivity, specificity, positive and negative predictive values of our model in the external cohort to determine the generalizability of the model. We conclude the chapter with a discussion on how the enhancement, sharing of resources, and collaboration between the BQC19 and Mount Sinai COVID-19 biobanks allows for the increase in the power and validity of the potential findings.

The novelty of the work presented in chapter 4 is in the inclusion of an independent cohort which represents best practices for model development and validation [46]. Testing of a prediction model in a cohort separate from the training cohort is vital. By including samples recruited from three separate hospitals, across two separate health care systems in two different countries, we increased the probability that the results presented are generalizable and not overfitted to the training data [47]. In addition, many previous studies that have tested the association between protein levels and COVID-19 outcomes have focused solely on circulating cytokines and chemokines [16, 48–54]. While this is a reasonable approach given the nature of the disease, we are unaware of any other studies of this scale that have tested the association of 4,701 circulating proteins with COVID-19 outcomes.

In summary, the overall structure for this thesis is as follows: chapter 2 contains pertinent background information required to understand the technicalities of the work presented in chapter 4 and includes a summary of existing literature on proteomics, genomics, and COVID-19. Chapter 3 provides a primer on machine learning and statistics related topics that are relevant to understanding the methods used in chapter 4. Chapter 4 describes the work performed on circulating proteins to identify a proteomic signature for COVID-19 outcomes. Finally, in chapter 5 we provide a conclusion to the thesis and discuss possibilities for future work.

Chapter 2

Background and Literature Review

In this section, we provide a review on the existing literature involving work on proteomics and genomics in the context of COVID-19 as well as subsections on relevant background knowledge important to later sections of this thesis.

Determining the genetic determinants of what causes severe COVID-19 disease is an important question. While it is already known that COVID-19 disproportionately affects older adults and that severe disease is more likely to occur in older adults [55] as well as those that are immunosuppressed [56], it is unknown why young, healthy individuals die from disease.

The predisposition for severe COVID-19 may lie within the genome and variation in the genetic variants everyone holds. By linking genomic data with high-dimensional proteomic data, novel findings related to COVID-19 infection can be rapidly discovered allowing for greater insights into the disease etiology.

The use of high-throughput proteomic analysis where thousands of protein measurements are extracted per sample has contributed insight into how plasma proteins may be used to assess and prevent future risk of common chronic or age-related morbidities. Williams et al. 2019 [18] have shown that protein scanning can serve as a source for observing potential health risks in individuals with only a single blood sample being required. In their work, they analyzed a similar set of 5,000 plasma proteins and developed predictive machine learning models to investigate the health state of individuals with regards to 11 different health indicators such as liver fat, kidney filtration, percentage body fat, visceral fat mass, lean body mass, cardiopulmonary fitness, physical activity, alcohol consumption, cigarette smoking, diabetes risk, and primary cardiovascular event risk. Their protein-phenotype models utilized elastic-net linear regression for predicting continuous outcomes, elastic net logistic regression for dichotomous outcomes, or survival models for longitudinal events. For dimensionality reduction, LASSO was used and false discovery rate (FDR)-corrected P values were used to rank candidate features. Models were trained using cross validation with data split into training, validation, and testing sets and evaluated using the AUC score, r^2, or accuracy. While effective in certain scenarios, such as the protein model for percentage body fat producing an r^2 of 0.91, the results produced by Williams derived data from a single blood draw with no follow up longitudinal analyses. In our case, data includes multiple sample points, allowing the assessment of the sensitivity of such longitudinal analyses. Furthermore, major features such as demographic data, risk factors, and genetic information were not included in their model development due to the goal of reducing costs and simplifying measures required for a future health checkup.

Other studies have also utilized protein-based models for predicting disease outcomes. Ganz et al. 2016 measured 1,130 proteins from 2,496 samples to predict cardiovascular outcomes [20]. LASSO was used as a form of feature selection to select 16 proteins and this set was further reduced down to a 9-protein model using stepwise backward elimination. The final 9-protein model was used to predict 4-year probability of myocardial infarction, stroke, heart failure, and all-cause death with performance measured by the C-statistic. In addition, plasma proteins have also been used as diagnostic tools in previous studies related to lung cancer [57,58] and a previous study implementing a 27-protein model for predicting cardiovascular disease risk has been shown to be an effective predictor of adverse effects of COVID-19 [19].

Previous studies have attempted to identify a proteomic signature for COVID-19 outcomes by profiling metabolic biomarkers using nuclear magnetic resonance spectroscopy [59], however this has limitations due to the intrinsic insensitivity of the methods. A team at the Massachusetts General Hospital (MGH) worked on a longitudinal COVID-19 study in collaboration with Olink Proteomics (a Swedish company specializing in proteomics and biomarker discovery) where over 1,400 plasma proteins were measured for each individual. The cohort involved 384 patients with 306 COVID positive and 78 virus-negative individuals who were enrolled due to the presentation of at-risk symptoms of COVID. Symptoms included ”1) tachypnea $\geq$ 22 breaths per minute; 2) oxygen saturation $\leq$ 92% on room air; 3) a requirement for supplemental oxygen; or 4) positive-pressure ventilation” [33]. Of the entire cohort, 306 patients (80%) tested positive for SARS-CoV-2 while the rest of the group (78 individuals) did not.

For this MGH cohort, over 1,400 plasma proteins were measured for each patient which was at the time of release, one of the largest longitudinal studies on COVID-19. Plasma protein levels were sampled from COVID positive patients (n = 306) over the course of a week with measurements on days 0, 3, and 7. For COVID negative individuals, plasma proteins were only sampled on day 0. In addition, COVID outcomes for the entire cohort were also collected and each individual had an outcome classified based on the World Health Organization (WHO) COVID-19 outcomes scale, which measures patient health status at defined time intervals over the course of a 28-day window once they have been diagnosed with COVID. The time points are at day 0, 3, 7, and 28 where for each day, individuals are classified using a numerical scale into one of the 6 categorical groups ranging from 1 to 6 with lower values representing higher severity of COVID infection. The scale is defined as follows: 1 = Death within 28 days; 2 = Intubated, ventilated, survived to 28 days; 3 = Non-invasive ventilation or high-flow nasal cannula; 4 = Hospitalized, supplementary O_2 required; 5 = Hospitalized, no supplementary O_2 required; 6 = Not hospitalized. In this study, protein levels were quantified according to

Normalized Protein eEXpression (NPX) levels where a larger magnitude indicates higher concentration of that protein.

While studies such as [33] have developed protein-based models for COVID outcomes, none have analyzed proteins at such a large-scale as performed in chapter 4 which contains work performed on one of the largest COVID-19 circulating protein datasets in the world. The identification of proteins that may be causally involved in the development of severe COVID-19 in patients is important and proteins such as OAS1 have been shown to affect COVID-19 outcomes [4]. Further, the gene coding for angiotensin-converting enzyme 2 (ACE2) which is a cell surface protein has been discovered to affect the ability of COVID-19 to enter host cells [60]. Thus, by analyzing massive scale proteomics and pinpointing important proteins involved in adverse COVID-19 outcomes, we can accelerate and enable drug repurposing, identification of new drug candidates, predict disease progression, and allow opportunities for further therapeutic development.

2.1 Basic Epidemiological Terminology

In this section, we provide an overview of a few key terms and important definitions that are relevant to later sections of this work.

- Prospective study (prospective cohort study): enrolling participants into study prior to the development of the disease or outcome. e.g. Randomized Controlled Trials (RCTs).

- Case control studies (retrospective studies, case-referrent studies): studies where the advantages include:

  - ability to simulataneously study multiple risk factors.
  - association between a risk factor and outcome can be initially tested for.
  - less time required than prospective studies due to disease or outcome already having happened.

- Case: a subject with a disease or outcome of interest.

- Control: a subject without the disease or outcome of interest present.

- Confounding: when a third variable is strongly associated with both the exposure and outcome variable and is not in the causal pathway between the exposure and outcome.

- Meta-analysis: combining data from multiple studies together to reach a conclusion. Advantages: greater statistical power, allows confirmation of results from individual studies, improved capability of generalizing, and extrapolating data to the general population affected by disease.

- Matched design: each case has a control that is matched (i.e. by age, gender) individually for that case. e.g. Case = 35 year old male with disease; Control: 35 year old male without disease.

- Unmatched design: controls are sampled from a non-affected population.

- Epidemiology: the study of health and disease patterns within the population. Distinction between classical epidemiology (without genetic factors) and genetic epidemiology (epidemiology with the use of genetic factors).

- disease etiology: the cause of a disease.

- Risk factor: usually the ”exposure variable”.

2.2 Human Genetics

The central dogma of molecular biology encompasses that shown in Figure 2.1. This diagram displays the flow of information from DNA to mRNA through the process of transcription and finally to proteins by a process called translation. Each step can be measured by different omics approaches such as genomics, transcriptomics, and proteomics and these technologies are important in investigating different biological questions. For instance, trait heritability can be studied through genomics. Proteins are the functional products within the body that are influenced by both an individual's genetics at birth and their environment and can be studied using proteomics. The effect of the environment can be at the micro level involving cells, or at the macro level involving diet, lifestyle changes, or pharmacological interventions involving drugs and chemicals. One of the central goals of human genetics is to identify genetic risk factors for diseases. Diseases may be complex diseases that are common within the population such as Alzheimer's disease, autoimmune diseases, type II diabetes, and many more. Other diseases such as sickle-cell anemia, Huntington's disease, and cystic fibrosis are examples of rare Mendelian diseases. A complex disease is one caused by a combination of factors including lifestyle choices, genetics, and the environment, while Mendelian disorders (single gene diseases) are caused by alterations in the genome of an individual and tend to be rare. Many tools exist for finding genetic risk factors and one such tool is a genome-wide association study (GWAS) which is described in section 2.3.

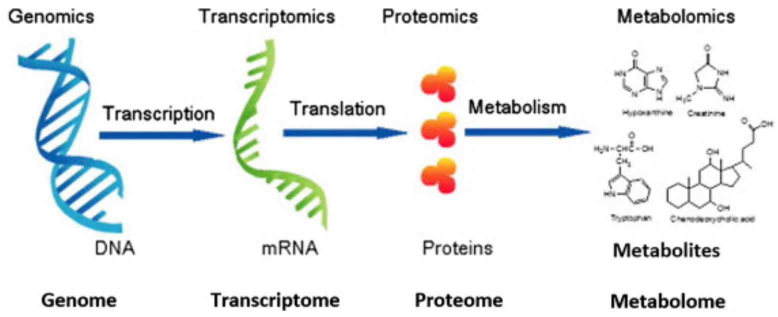


Figure 2.1: OMICS technologies: Moving from DNA to phenotype. Retrieved from [61]

2.3 Genome-Wide Association Studies

Genome-wide association studies (GWAS) provide a way for investigating the genetic architecture of common diseases [62] and also relatively rare diseases [63]. Genetic variation in the genome occurs through single nucleotide variants (SNVs) which are any single nucleotide change that occurs in the genome (rare and common single nucleotide variations) or indels which are insertions or deletions of bases that occur in an organism's genome. In DNA, there are four nucleotides: guanine (G), adenine (A), cytosine (C), and thymine (T). Thus, an example of a SNV can be a replacement of the nucleotide thymine (T) with the nucleotide cytosine (C) within the genome. Single nucleotide variants (SNVs) and indels are one of the most common types of genetic variation which are identified by GWAS to determine how genetic variations contribute to an individual's risk for complex diseases. Some SNVs are known as single nucleotide polymorphisms (SNPs) if the variant is a common variant present in 1-5% of the population. SNPs will rarely have a high impact or effect on the disease or gene function whereas less common SNVs and indels (rarity due to selective pressure) may have real high effect on the function of a gene. For example, when the position of a SNV is within a gene or in close proximity to a gene such as in a regulatory region, they may act as important biomarkers enabling the elucidation of gene-disease associations and may even affect gene function.

The association between genetic variation and diseases can be investigated through GWAS and visualized with Manhattan plots (Figure 2.2). Manhattan plots display the association between genetic variants in the entire genome (X-axis) and a disease outcome, for example COVID-19 severity, represented by the statistical significance or P value (Y-axis) in negative logarithmic scale. Each data point represents a single SNP or indel and displays the association between that SNP at a particular position on a chromosome and a trait of interest across all individuals.

Statistical significance in GWAS requires a P value less than 5×10^{-8} due to there being approximately 1 million common, independent SNPs in the genome ($\frac{0.05}{1 \times 10^{-6}}$). Note that there are different terms for SNPs such as "tag" SNPs which are those used on genotype microarrays and "lead" or "index" SNPs which refer to SNPs that have the smallest and therefore most significant P values in a particular region of interest [64].

GWAS have many applications and one of the ways they can be used is to generate polygenic risk scores (PRS). A polygenic risk score is a single scalar value estimating the genetic risk of an individual for a phenotypic trait or disease. The calculation of a PRS involves summing over all of an individual's risk alleles and weighting each risk allele by the effect size computed for that allele from GWAS. Note that in the construction of a PRS, linkage disequilibrium (LD) must be accounted for when computing weights and this is done by linkage disequilibrium pruning, which removes loci that have a large pairwise LD value. By providing a single value estimate of an individual's genetic liability to a trait or disease, polygenic risk scores combine and summarize the effects of many SNPs and enable the ability to provide a risk prediction of complex traits such as mental disorders, autoimmune diseases, cardiovascular disease, diabetes, height, and hair colour. In the future, polygenic risk scores may be more widespread in precision and personalized medicine for the prediction and prevention of disease.

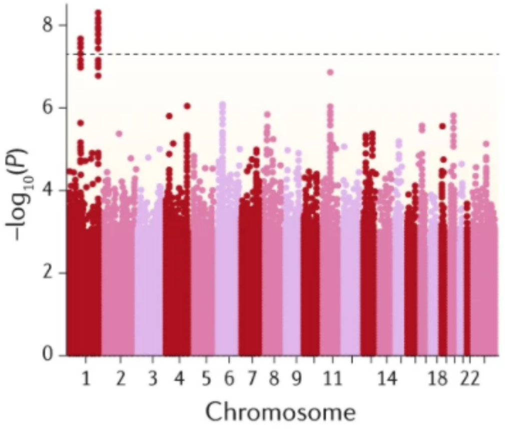


Figure 2.2: Manhattan plot. Retrieved from [64]

2.4 Multiple Testing

The problem of multiple comparisons or multiple testing occurs when many statistical inferences are made simultaneously with each statistical test likely to yield a discovery. For example, if 1 million SNPs are tested for association at the 5% level of significance, the expected number of false positives would be 50,000 [65]. In genomics, there are two common methods for addressing the multiple comparisons problem which are discussed in this section.

2.4.1 Bonferroni Correction: Controlling the Familywise Error Rate

The Bonferroni correction is a more conservative approach and one of the simplest and most widely used methods for counteracting the problem of multiple comparisons [66]. The occurrence of false positives or Type I errors is controlled by thresholding the familywise error rate or the probability that a false conclusion is made when a series of hypothesis tests are performed.

$$Bonf(\mathrm{P}) = \mathrm{P}/i$$

where i is the number of tests.

For example, if 1,000 *independent* tests are performed and the threshold for significance has been set to P<0.05 then the Bonferroni corrected P value to be significant is P$< \frac{0.05}{1000} = 5 \times 10^{-5}$.

Generally, when the entire set of tests performed are all independent from one another, then the adjustments made by the Bonferroni correction method will be perfect. However, if positive correlation occurs between any of the test statistics (and thus some tests are dependent), the Bonferroni correction may not be the most suitable method in multiple hypothesis testing. Despite this fact, the Bonferroni correction is still widely accepted and used in genetic association studies. A reason being due to high false positivity rates in GWAS due to population structure and family relatedness which is mitigated by being overly conservative. However, due to the conservative nature of this approach, a consequence is the reduction of statistical power resulting from the increased probability of producing false negatives. Thus, other methods should be used such as the Benjamini-Hochberg procedure [67] to prevent the loss of potentially valuable information.

2.4.2 Benjamini-Hochberg Procedure: Controlling the False Discovery Rate (FDR)

Aside from the Bonferroni correction, an alternative method for controlling for multiple comparisons is the Benjamini-Hochberg (BH) procedure [67] which is more practical and widely used in genomics problems.

If the goal of a study is discovery, the BH procedure can be used and is preferable to the Bonferroni correction method. The idea of discovery in a statistical context constitutes the rejection of a hypothesis (e.g., reject the null hypothesis. In a GWAS context, the null hypthesis is that none of the SNPs are associated with the disease or trait of interest). Therefore, controlling for the false discovery rate involves limiting or setting a threshold for the number of incorrect rejections of a hypothesis. The false discovery rate is the likelihood that an incorrect reject of a hypothesis occurs and can be preset according to the problem of interest. For example, setting an FDR of 10% corresponds to a willingness to have 10% of results being false positives.

$$p_{(i)}^{BH} = \min\{\min_{j \geq i}\{\frac{mp_{(j)}}{j}\}, 1\}.$$

where m is the total number of tests, j is the rank order of the P value, and p is the P value. The method is as follows:

1. First, order all P values from small to large. Then multiply each value by the total number of tests and divide by its rank order.

2. Second, make sure that the resulting sequence of P values is non-decreasing: if it ever starts decreasing, make the preceding value equal to the subsequent (repeat, until the whole sequence becomes non-decreasing).

3. If any value ends up larger than 1, make it equal to 1.

Chapter 3

Primer on Machine Learning

Machine Learning (ML) has been commonly defined as the set of algorithms and statistical models that are used by computer systems to perform a specific task without the use of explicit instructions. ″A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E″ [68]. In essence, the goal is to build a model that best represents the structure of a given dataset and use this model to predict or generalize to unseen data.

A machine learning model is given some input (also known as variables, **features**, attributes, predictors) and produces an output (also known as variables, targets, **labels**, predictions). To train such a model, the dataset is generally divided into a training set and a testing set. The training set is used to build and update the model and evaluation of the final performance or generalizability of the model is performed by running the trained model on the test set [69]. The algorithm should not have access to the test set during training to prevent information leakage and prediction bias. Since the test set is used to evaluate machine learning methods, the performance of a model on testing data is the most important indicator of performance.

There are three major recognized categories in machine learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning trains models on data where the desired outcome or labels are already known and includes two categories of algorithms: regression and classification. In classification, the predictions are categorical variables representing discrete values, whereas in regression, the output produced by the model are numerical and represent continuous values. Unsupervised learning involves training models with unlabeled data. This category of machine learning involves methods such as *clustering*, where the labels or classes are inferred and not provided to the algorithm, *dimensionality reduction*, and *generative modeling*, where the algorithm models the distribution of data and learns to generate the data instead of directly categorizing the instances into different classes. Reinforcement learning involves an algorithm known as an “agent” that interacts with an environment, and uses weak supervision through the reward signal to perform tasks within the environment. Semi-supervised machine learning algorithms also exist. These algorithms combine aspects from both supervised and unsupervised learning and uses a few labeled examples to train itself.

While machine learning has been used in a variety of applications and proven effective in domains such as healthcare [70], image recognition [71], self-driving cars [72], fraud detection [73] and many more, there are many shortcomings that still need to be addressed involving privacy, bias (errors due to incorrect assumptions of the model), generalization [74], and the misuse of machine learning algorithms such as deepfakes [75].

In the following subsections, linear models are explored in more detail as they are discussed significantly in chapter 4 of this thesis.

3.1 Regression Models

In this section, we provide background on types of generalized linear models used in later sections of this thesis. Generalized linear models are a generalization of the concepts and abilities of regular linear models, and models such as linear regression and logistic regression are detailed in the following subsections.

3.1.1 Linear Regression

Linear regression also referred to as least squares regression attempts to discover the best fitting line to the data [76]. The goal is to find the line that results in the minimum sum of squared residuals where the residuals are the differences between the real data and the predicted value by the model. We define x as the predictor variable (also termed the input or independent variable) and y as the response variable (also termed the output or dependent variable). Assuming we have $i = 1, \ldots, n$ instances, we want to minimize $\sum_{i=1}^{n} r_i^2$, where r_i is the residual defined as $r_i = y_i - f(x_i, \beta)$ and β is the vector of weights. The f is the model or function which takes an input x and outputs a prediction y. In two dimensions, $f(x_i, \beta) = \beta_0 + \beta_1 x$ which is a straight line. The ultimate goal is to find the optimal values for the intercept β_0 and the slope β_1. Linear regression is also termed "Least squares" since the final model finds the line that results in the smallest sum of squares by searching for values for β_i which provide the best fitting line to the training data.

If we define $\mathbf{X}$ as a matrix with full column rank, then the analytical solution for a linear regression model of the form $\mathbf{y} = \mathbf{X}\beta$ has the least squares solution

$$\hat{\beta} = \arg\min \|\mathbf{X}\beta - y\|_2,$$

which is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

However, if $\mathbf{X}$ is large and dense, and the closed form solution becomes computationally intractable, an iterative algorithm like gradient descent should be used to fit the model parameters [77].

Theoretically, gradient descent is an algorithm that minimizes functions. Given a function defined by a set of parameters, gradient descent starts with an initial set of parameters and iteratively moves towards a set of parameter values that minimize the function.

For the iterative minimization step, we take steps in the negative direction of the function gradient.

Given Error function: $Error_{(m,b)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$, we compute a partial derivative for each parameter (in this case, m and b)

$$\frac{\partial}{\partial m} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial b} = \frac{2}{N}\sum_{i=1}^{N} -(y_i - (mx_i + b))$$

where m is defined as the slope or gradient of the function and b is the bias parameter. In the algorithm, we can initialize our search at any pair m and b values. Each iteration will update m and b to give a line that gives slightly less error than the previous iteration and thus gives a better fitted line to the data.

In linear regression, there is a single global minimum so the error surface is convex [78]. If a problem has local minima, then gradient descent may get stuck and the solution is to use stochastic gradient descent [79] to solve this optimization problem. To determine when to stop the descent, we look for small changes in error iteration-to-iteration to determine convergence.

In general, in most nonlinear regression problems, such as logistic regression discussed in section 3.1.2, no closed form solution exists and numerical or iterative methods must be used to solve the optimization problem.

Linear Mixed Models

Fixed effects model are models where model parameters are fixed or non-random, while random effects models, also called Variance components models, are models where the parameters are random variables [80]. In a Mixed model [81] (also known as Mixed-Effects Model or Mixed Error-Component model) both fixed effects and random effects are included. A Linear Mixed model can be defined as

$$y = X\beta + Zu + \epsilon,$$

where y is the target vector, X is the fixed effect features (unknown), β is the estimate of the fixed effect, u is an unknown vector of random effects, ϵ is an unknown vector of random errors, and Z is the random effect features (unknown).

3.1.2 Logistic Regression

Odds

The odds represent the likelihood of a particular outcome. Odds do not characterize probabilities but instead represent the ratio of an event happening to an event not happening. For example, $\frac{\text{disease}}{\text{no disease}}$. Odds can be computed from probabilities: $\text{Odds}(\text{disease}) = \frac{p(disease)}{1-p(disease)}$ or from counts: $\frac{\text{number of times disease occurs}}{\text{number of times no disease occurs}}$

Log Odds

Assume that the odds of having a disease is 1 to 4, $\frac{1}{4} = 0.25$. If the odds of having this disease was worse, for example 1 to 32, $\frac{1}{32} = 0.031$ then we notice that as the odds of having a disease become worse, the odds approach closer and closer to 0. In other words, if the odds were against having a disease, then they will lie between 0 and 1.

On the other hand, assume that the odds of having a disease were high, for example 4/3, or 1.3. If the odds of having this disease were even higher such as 32/3 or 10.7 then we notice that the more likely the disease, the higher the odds and this value can increase without bound. Thus, the odds of having this disease will range between 1 and ∞.

Due to the large asymmetry between the odds for ($\text{Odds} \in [1, \infty)$) and against ($\text{Odds} \in [0, 1]$) a disease, this creates difficulty in comparing them. By taking the $\log$ of the odds, this incompatibility can be mitigated since the log scale is symmetrical and easier to interpret. Thus, if the odds are against a disease, for example, 1 to 6 then the $\log(\text{Odds})$ are

$\log(1/6) = \log(0.17) = -1.79$ and if the odds are for a disease 6 to 1, then the $\log(\text{Odds})$ is $\log(6/1) = 1.79$, which is symmetrical and comparable.

For logistic regression, we can make the model assumption:

$$\log \frac{p(y|x_1, x_2)}{1 - p(y|x_1, x_2)} = \text{model}(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \text{ (linear model)} \tag{3.1}$$

$$y \in \{0, 1\}, x_1, x_2 \in \mathbb{R}, \tag{3.2}$$

where $log(\frac{p}{1-p})$, the log of the ratio of the probabilities is called the logit function and forms the basis for logistic regression.

For a given input x (commonly termed the exposure), the odds are defined by

$$\text{Odds}(x) = \frac{p(1|x_1, x_2)}{p(0|x_1, x_2)} = \frac{p(1|x_1, x_2)}{1 - p(1|x_1, x_2)} = \exp(\text{model}(x_1, x_2)). \tag{3.3}$$

For given $(x_1, x_2), (x_1', x_2')$, the odds ratio can be defined as follows:

$$\frac{\text{Odds}(x_1', x_2')}{\text{Odds}(x_1, x_2)}.$$

The odds ratio refers to a ”ratio of odds”, and calculates the ratio between the odds of the outcome in the presence of the exposure, and the odds of the outcome in the absence of the exposure. The values correspond to effect sizes where larger values mean that the exposure is a good predictor of the outcome, while smaller values mean that the exposure is not a good predictor of the outcome.

Using the odds ratio, we can solve the logistic regression model:

$$\begin{aligned}
&\log \frac{\mathrm{Odds}(x_1+1, x_2)}{\mathrm{Odds}(x_1, x_2)} \\
&= \log \mathrm{Odds}(x_1+1, x_2) - \log \mathrm{Odds}(x_1, x_2) \\
&= (\beta_0 + \beta_1(x_1+1) + \beta_2 x_2) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\
&= \beta_1, \\
&\frac{\mathrm{Odds}(x_1+1, x_2)}{\mathrm{Odds}(x_1, x_2)} = e^{\beta_1} \\
&\frac{\mathrm{Odds}(x_1, x_2+1)}{\mathrm{Odds}(x_1, x_2)} = e^{\beta_2}.
\end{aligned}$$

Logistic Regression is a specific type of Generalized Linear Model (GLM) where the outcome depends on the sum of the inputs and parameters. The model can make predictions whether the dependent variable is dichotomous or contains multiple classes. The classifier uses a logistic function and assumes the absence of any outliers which are removed after converting continuous features to standardized scores. Logistic regression is similar to linear regression except it is used for classification and predicts categorical instead of continuous outcomes. Thus, instead of fitting a line to the data, logistic regression fits an ″S″-shaped ″logistic function″ (refer to Figure 3.1). The range of the logistic function lies between 0 and 1 and represents the probability that the independent variable X is equal to the dependent variable y. This function allows the modelling of binary outcomes using the sigmoid function:

$$\sigma(t) = \frac{1}{1+e^{-t}}.$$

Logistic regression is generally a good option as a classifier when features are correlated with one another. However, the correlation between features should be less than 0.90, and as long as this is satisfied, the assumption of absence of multicollinearity is met [83]. Moreover, logistic regression is a model that is used to calculate the probability

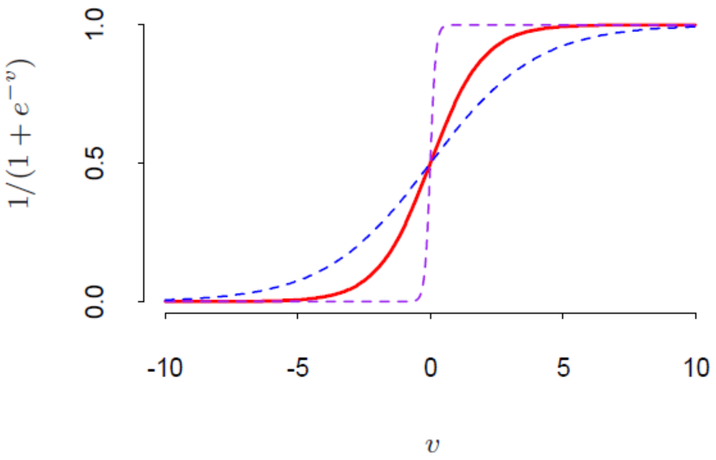


Figure 3.1: Plot of logistic function. Retrieved from [82].

of a model belonging to a class and uses a maximum a posteriori estimation, instead of directly predicting by classification [84]. If some features are correlated with each other, logistic regression is affected to a lesser degree as it compensates by weighting the features lower. However, if unnecessary features are not removed, logistic regression may make less accurate predictions. Furthermore, when training size is limited, logistic regression may overfit because of insufficient samples. When there are fewer training instances than features, Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression can serve as regularization methods which may help improve the performance of logistic regression. Both of these methods are explained in further detail in sections 3.3.1 and 3.3.2, respectively.

3.2 Bias-Variance Trade-off

It is important to note that training of models such as logistic regression requires optimizing a tradeoff between bias and variance, often termed the Bias-Variance Trade-off. Bias is the inability for an ML algorithm to capture the true structure of the data due to incorrect assumptions of the model. For example, if data points in a 2D space were arranged in a curve but the model fitted to this data is a straight line, this represents high bias. A case of low bias would occur when the model fits every single data point in the training set. Variance occurs when the predictions of the model fluctuate when small changes to the data are made. For example, a model that is overfitting the training data would result in the case of high variance. If a model fits the training set well but not the testing set, then we say the model is overfit. In machine learning, the ideal algorithm has low bias (signifying that the model is accurately modelling the true relationship within the data) and low variance (suggesting that the model is producing consistent predictions across different datasets). This is accomplished by searching for an optimal choice between a simple model and a complex model. There are three commonly used methods for finding this ideal solution between simple and complicated models which are regularization, boosting, and bagging. Regularization is one of the main methods to avoid overfitting which reduces the generalization and accuracy of a model and is used in the work presented in chapter 4. The following section explains regularization in more detail.

3.3 Sparse Models and Regularization

Regularization is a technique for shrinking and constraining model coefficients to reduce the complexity of the model and thus reduce the possibility of overfitting to the training set during model training. There are three regularization techniques that are commonly used to decrease the size of coefficients which are discussed in sections 3.3.1, 3.3.2, 3.3.3: L_1 regularization (LASSO Regression) [85], L_2 regularization (Ridge Regression) [86], and Elastic Net which uses a linear combination of L_1 and L_2 penalties from Lasso and Ridge regression [87].

3.3.1 L1 Regularization

The Least Absolute Shrinkage and Selection Operator (LASSO) regression method penalizes the L_1 norm by regularizing on the absolute sum of the model coefficients. Logistic

Regression with L1 regularizer (LASSO) minimizes the cost function:

$$\min_{w,b} ||w||_1 + \frac{1}{\lambda}\sum_{t=1}^{n} \log(e^{(-y_i(X_i^T w+b)+1)}), \tag{3.4}$$

where a small λ corresponds to less regularization. Note that notation used here assumes binary classification, $y_i \in \{-1, 1\}$.

Since LASSO induces sparsity within the model parameters (w in Equation 3.4) it can be used as a form of feature selection. Feature selection is used to avoid overfitting and retain the predictive power of the model. This is especially important in cases where features outnumber the number of observations (samples). As the penalty term λ increases, LASSO sets more coefficients of features to zero which reduces the number of dimensions. However, a limitation of LASSO is that if two features are collinear, it randomly selects one of them.

3.3.2 L2 Regularization

In cases where features are highly correlated (multicollinearity), ridge regression can be used which penalizes coefficients using L_2 regularization.

Logistic Regression with L_2 regularizer (Ridge) is defined as follows:

$$\min_{w,b} \frac{1}{2}w^T w + \frac{1}{\lambda}\sum_{t=1}^{n} \log(e^{(-y_i(X_i^T w+b)+1)}), \tag{3.5}$$

where a small λ corresponds to less regularization. Note that notation used here assumes binary classification, $y_i \in \{-1, 1\}$.

The error function can equivalently be written as:

$$Error(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}||w||^2,$$

(usually divide Least Squares by N to scale the error) where $||w||^2 \equiv w^T w = w_0^2 + w_1^2 + \ldots + w_M^2$ i.e. the square of the slope, which we denote as Δ^2.

Note that Δ^2 (usually the bias term w_0, sometimes denoted by b, or intercept term, is excluded from the regularizer) adds a penalty to the traditional Least Squares method, while the λ term determines how severe that penalty is.

Ridge regression shrinks coefficients but never sets them to zero. Thus, it does not reduce the number of features since none of the features have coefficients set to 0. The main idea behind Ridge Regression is to reduce the complexity of the model in order to prevent the model from fitting the training data too closely. In other words, L_2 regularization introduces a small amount of bias into the model and in return results in a significant drop in variance. By beginning with a slightly worse fit, Ridge Regression can provide better long term predictions.

To determine the choice of λ, many different values must be tested. A method for finding the optimal λ is by cross-validation (section 3.4), which determines which λ parameter results in the lowest expected error. Applying L_2 regularization to Logistic Regression involves optimizing the sum of the likelihoods instead of the squared residuals such as in linear regression since logistic regression is solved using Maximum Likelihood. i.e. sum of likelihoods + $\lambda \times \Delta^2$ (excluding the bias term in the slope)

3.3.3 Elastic Net Regularization

Elastic net regularization is a generalized form of LASSO and Ridge and combines both L_1 and L_2 regularization in its penalty term. Logistic Regression with elastic-net regularizer is defined as:

$$\min_{w,b} \frac{1-\alpha}{2} w^T w + \alpha ||w||_1 + \frac{1}{\lambda} \sum_{t=1}^{n} log(e^{(-y_i(X_i^T w + b) + 1)}), \tag{3.6}$$

where α determines the ratio of L1 vs. L2 regularization, and a small λ corresponds to less regularization. Note that notation used here assumes binary classification, $y_i \in \{-1, 1\}$.

Elastic net is particularly useful when correlation occurs between multiple features in the data (LASSO selects one randomly while elastic-net will select both). Due to this reason, in general, elastic net performs better than LASSO on data with many correlated features.

3.4 Cross-Validation

Regularization is important due to its ability to reduce the variance of the model without increasing the bias substantially. However, the value of the tuning parameter λ must be carefully chosen. Larger values of λ reduce the size of coefficients and thus reduce the model variance but as λ increases more, the reduction in coefficient sizes leads to important information being lost and results in an increase in the bias. One way to tune the hyperparameters within a model is cross-validation.

Cross-validation provides a few advantages aside from selection of optimal hyperparameters such as allowing the comparison of different learning procedures (comparing different algorithm performances) and in the case where data is limited, cross-validation allows for the better usage of all data. Further, it can provide a preliminary estimate of the generalization performance of the model as well as provide statistics such as the mean and variance during training.

Cross-validation can be generalized as N-fold cross-validation where data is split into N blocks and 1 block is used for testing. An extreme case of cross-validation is Leave One Out Cross-Validation [88] where each data point is a block so each data point is used as the validation set eventually. The choice of N in N-fold cross-validation generally depends on the size of the dataset as well as the computational budget. An illustration of 5 sets of five-fold cross-validation is provided in figure 3.2 which splits data into folds and uses each fold for validation and the rest for training.

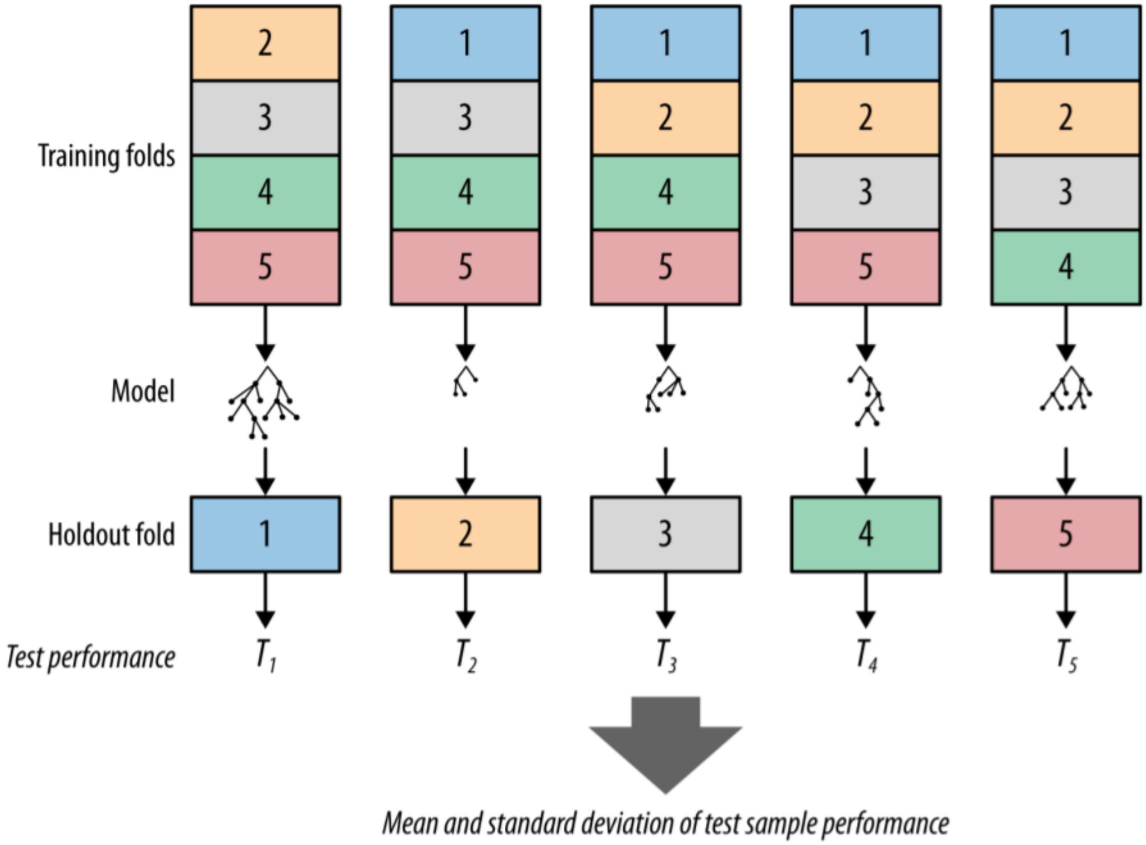


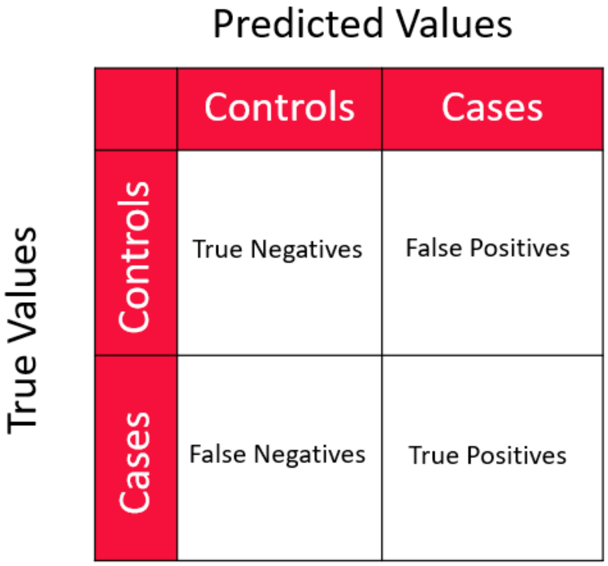
Figure 3.2: Illustration of Cross-Validation. Retrieved from [89]

3.5 Model Evaluation and Assessment

In this section, we provide a few metrics for determining and evaluating the performance of a model.

3.5.1 Confusion Matrix

The confusion matrix provides a table layout to determine the performance of a machine learning algorithm [90]. An example is shown in figure 3.3 where each position in the matrix holds an integer value for the number of cases classified into each positive or negative category. Columns are shown as the prediction of the algorithm and rows are the ground truths. A prediction is a true positive (TP) when an algorithm correctly predicts the presence of a characteristic (such as the presence of a disease). When an algorithm correctly predicts the absence of a characteristic (absence of a disease), this is known as a true negative (TN). On the other hand, if an algrithm predicted the presence of a characteristic when the characteristic was absent this is known as a false positive (FP). Predicting the absence of a characteristic when the characteristic is in fact present is known as a false negative (FN). The true positives (TP) and true negatives (TP) make up the number of correct predictions while the false positives (FP) and false negatives (FN) represent the number of times the algorithm outputted an incorrect prediction. Generally, a model should be selected which produces the highest number of correct predictions (TP+TN) and lowest number of incorrect predictions (FP+FN). Additional metrics can also be used to evaluate a model such as the sensitivity, specificity, and area under the curve (AUROC) which are described in the following sections.



		Predicted Values	
		Controls	Cases
True Values	Controls	True Negatives	False Positives
	Cases	False Negatives	True Positives

Figure 3.3: Example of a confusion matrix.

3.5.2 Precision/Recall Curves; Sensitivity/Specificity

The confusion matrix allows the calculation of more performance metrics to evaluate the predictive abilities of a model. A few metrics that are used in chapter 4 are discussed here.

One such metric is the sensitivity (also known as the true positive rate) which is the probability of a positive prediction.

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{\text{number of people with illness}}.$$

Another metric is the specificity (also known as the true negative rate) which represents the probability of a negative prediction.

$$\text{Specificity} = \text{Selectivity} = \frac{TN}{TN + FP} = \frac{TN}{\text{number of people without illness}},$$

$$\text{Precision} = \text{Positive Predictive Value} = \frac{TP}{TP + FP},$$

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}.$$

3.5.3 Area under the Receiver Operating Characteristic

The area under the receiver operating characteristic (AUROC) or area under the curve (AUC) provides a way of measuring the performance of classification algorithms [91]. It can be used to measure the goodness of fit for logistic regression binary outcomes. An example is shown in figure 3.4. The Y-axis displays the True Positive Rate (Sensitivity) while the X-axis displays the False Positive Rate (1 - Specificity). The AUC is also commonly referred to as the C-Statistic (concordance statistic, C-index) [92].

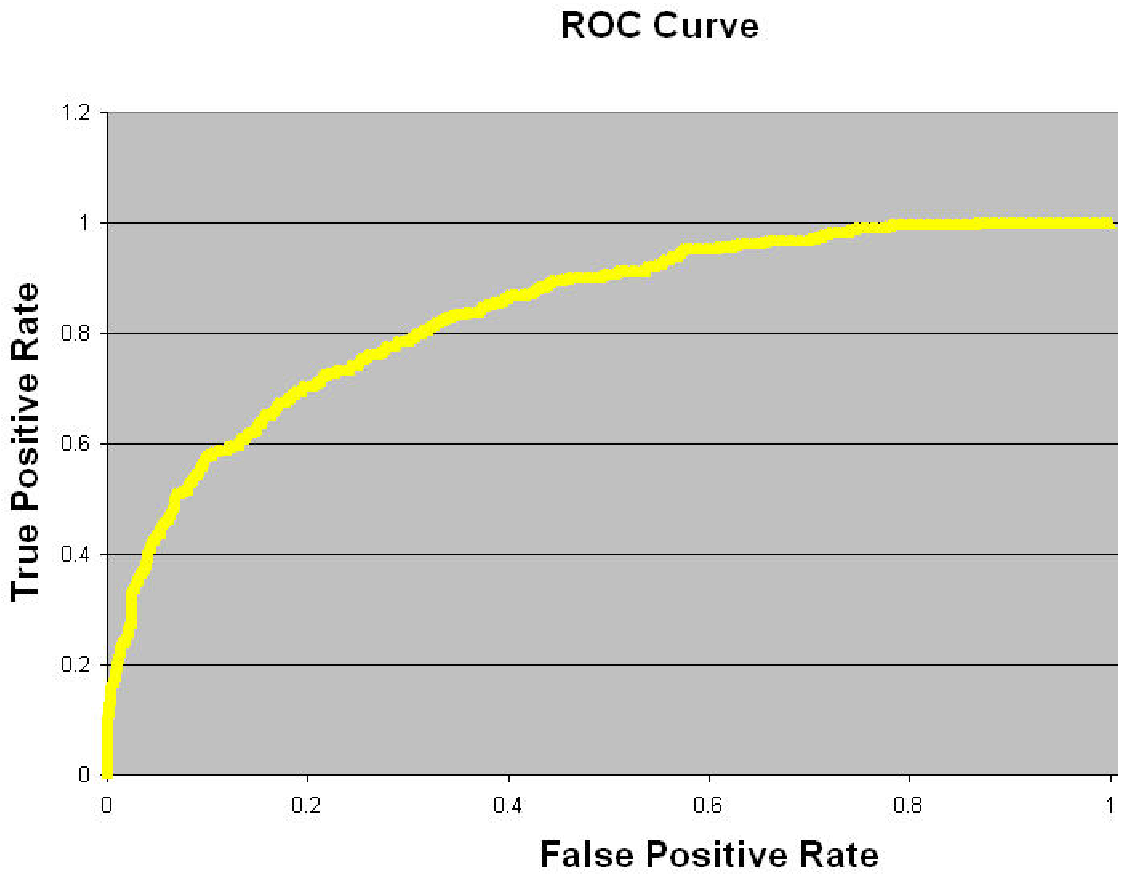


Figure 3.4: Area Under the Receiving Operating Curve example. Retrieved from [93].

3.5.4 Youden's J Statistic

Youden's J statistic is a metric often used with AUROC analysis that allows the evaluation of the performance of binary classifiers [94]. It is defined as follows:

$$\begin{aligned} J &= \text{Sensitivity} + \text{Specificity} - 1, \\ &= \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1. \end{aligned}$$

Youden's J ranges between 0 and 1 (inclusive) where a perfect model would produce a value of 1 representing the absence of false positives or false negatives.

The statistic can be computed at every single point of an AUROC curve to determine the optimal threshold for the outcome prediction. By selecting the point on the AUROC curve corresponding to the largest J, the sensitivity and specificity of the classifier can be maximized since this point corresponds to the threshold the model should use to output dichotomous predictions.

3.6 Dimensionality Reduction

Dimensionality reduction involves the transformation of high dimensional data to a low dimensional space while preserving the intrinsic properties of the original data. Dimensionality reduction can be used for clustering, reducing noise in the data, and for data visualization.

Both linear and nonlinear methods exist. One of the most extensively used linear methods is Principal Component Analysis (PCA) [95]. In essence, principal component analysis is a linear dimensionality reduction method that transforms data within a high dimensional space and applies linear transformations to it to reduce it down to a lower dimensional space while minimizing information loss. It is useful in dimensionality reduction, data visualization, and feature extraction and essentially displays highly correlated data close to one another in a cluster. Importantly, the principal component axes are ranked in order of the percentage of variability they explain. Since principal component 1 (PC1) explains more variability than PC2, if two clusters are the same Euclidean distance on the PC1 axis compared to two clusters that are the same distance along the PC2 axis, then the two clusters along the PC1 axis are more different from one another than the two clusters on the PC2 axis.

For nonlinear methods, Uniform Manifold Approximation and Projection (UMAP) [96] is commonly used which improves on the earlier method t-distributed Stochastic Neighbor Embedding (t-SNE) [97] in speed and preservation of global structure. UMAP makes assumptions about the data that allows for the lower dimensional embedding to be modelled as a fuzzy topological structure. This fuzzy complex represents a weighted graph with edges being the likelihood of two points being connected to one another. UMAP extends a radius outward from each point and assumes points are connected if their radii intersect. As the radius increases, UMAP decreases the likelihood that points are connected thereby creating a ”fuzzy” structure. Compared to PCA, UMAP may achieve better data separation and is more effective at displaying clusters of data and their proximity to one another. PCA, on the other hand, may be highly influenced by outliers in the data and is also unable capture nonlinear dependencies due to being a linear projection.

Chapter 4

Circulating Proteins to Predict Adverse COVID-19 Outcomes

In this chapter, we describe how circulating proteins can be used to predict adverse COVID-19 outcomes. We undertook a large-scale study to assess the relationship of thousands of circulating proteins with COVID-19 outcomes. To do so, we used machine learning methods to develop a predictive model of COVID-19 severity using the circulating blood protein abundances as predictors. Proteins were measured using 4,984 nucleic acid aptamers (SOMAmer reagents) [44] targeting 4,701 unique circulating human proteins in two cohorts collected from two countries, which in total included 986 individuals. The training cohort was comprised of 417 individuals from the Biobanque Québécoise de la COVID-19 (BQC19 cohort). This cohort was used to train a model to predict adverse COVID-19 outcomes. This model was then tested in a separate test cohort of 569 individuals from the Mount Sinai Hospital in New York City, which was similarly characterized for the same protein measurements and COVID-19 outcomes.

This large-scale study across two countries and two geographically separated cohorts identified circulating proteins associated with COVID-19 outcomes measured at a large-scale in well-characterized cohorts. These findings provide insights into the biological pathways influencing these outcomes and the ability of proteomics to predict these outcomes.

4.1 Methods

In this section, we describe relevant information pertaining to the conduction of the project as well as technical details of specific experimental procedures, datasets, and selection of models.

4.1.1 Cohorts

The Biobanque Québécoise de la COVID-19 (BQC19) is a Québec-wide biobank which was launched to enable research into the causes and consequences of COVID-19 disease (see bqc19.ca) [42]. It includes datasets on clinical and phenotypic data, transcriptomic data, immune-serology, and metabolomic data. The BQC19 recruits patients from around Quebec, Canada and collects blood samples from participants presenting severe and non-severe COVID-19 to enable data sharing among researchers around the world in a global effort to determine the etiology behind SARS-CoV-2 infection. In addition, the BQC19 also contains control samples which are collected from participants without COVID-19. Support for the BQC19 is provided by the Fonds de recherche du Québec – Santé (FRQS) and Génome Québec and the Public Health Agency of Canada. To date (as of Dec 10, 2021), blood samples have been collected from 3,553 consenting Quebecers across four regions of the province at nine different clinical sites totaling 30,353 samples.

For our study, we used BQC19 blood samples from 417 patients (313 SARS-CoV-2 nasal swab PCR positive patients (mean and median time since symptom onset in COVID-19 patients = 7.0 days, respectively; standard deviation = 3.96 days) and 104 individuals who presented with symptoms consistent with COVID-19 but had negative SARS-CoV-2 PCR nasal swabs) with available proteomic data from the SomaScan SomaLogic® assay. The subjects were recruited at the Jewish General Hospital (JGH) and Centre Hospitalier de l'Université de Montréal (CHUM) in Montréal, Québec, Canada, both of which are university affiliated hospitals. For each individual, blood samples drawn at the earliest time point were used for training when an individual had multiple blood draws. Selecting the blood sample at the earliest time point reflects the protein measurements during the acute phase of COVID-19 disease. The demographic characteristics of the participants in the BQC19 cohort who underwent SomaScan® assays is detailed in Table 4.1. The demographic characteristics were obtained by medical chart review or patient interview performed by trained clinicians or trained research coordinators.

The Mount Sinai cohort used in this study was composed of blood samples collected from 569 patients made up of 472 SARS-CoV-2 positive patients and 89 SARS-CoV-2 negative patients confirmed through PCR tests, one COVID-19 positive patient diagnosed by a chest CT while the remaining 7 individuals were COVID-19 negative and did not have COVID-19 symptoms during specimen collection but may have had a history of exposure. The samples donated by the patients in the Mount Sinai cohort underwent the same proteomic data collection and profiling performed as in the BQC19 cohort. The subjects were recruited at the Mount Sinai Hospital in New York City which is affiliated with the Icahn School of Medicine. Table 4.1 lists the demographic and sample processing parameters of participants in the Mount Sinai cohort that underwent SomaScan® assays. Demographic characteristics were obtained similarly to that of the BQC19 cohort.

4.1.2 Demographic, Sample Processing, and Clinical Variable Definitions

Age and sex from the BQC19 and Mount Sinai cohorts were collected. Sample processing time and hospital site were collected for BQC19 samples with the former being a continuous variable that quantifies the time in hours between sample collection and sample freezing.

Dataset	**Training Cohort (BQC19)**				
COVID-19 Severity		Severe		Critical	
	All Samples (n=417)	Cases (n=175)	Controls (n=242)	Cases (n=93)	Controls (n=324)
Age in years *	65.3 (18.4)	67.6 (17.6)	63.7 (18.8)	66.7 (16.5)	64.9 (18.9)
Female sex **	200 (48.0)	82 (46.9)	118 (48.8)	36 (38.7)	164 (50.6)
Sample Processing Time (hours) *	9.7 (13.4)	7.5 (6.5)	11.3 (16.5)	7.0 (6.3)	10.4 (14.8)
Hospital Site **					
CHUM	105 (25.2)	42 (24.0)	63 (26.0)	32 (34.4)	73 (22.5)
JGH	312 (74.8)	133 (76.0)	179 (74.0)	61 (65.6)	251 (77.5)
Diabetes **					
No	288 (69.1)	100 (57.1)	188 (77.7)	54 (58.1)	234 (72.2)
Yes	127 (30.5)	75 (42.9)	52 (21.5)	39 (41.9)	88 (27.2)
Chronic Obstructive Pulmonary Disease **					
No	358 (85.9)	145 (82.9)	213 (88.0)	77 (82.8)	281 (86.7)
Yes	57 (13.7)	30 (17.1)	27 (11.2)	16 (17.2)	41 (12.7)
Chronic Kidney Disease **					
No	363 (87.1)	151 (86.3)	212 (87.6)	77 (82.8)	286 (88.3)
Yes	52 (12.5)	24 (13.7)	28 (11.6)	16 (17.2)	36 (11.1)
Congestive Heart Failure **					
No	355 (85.1)	152 (86.9)	203 (83.9)	80 (86.0)	275 (84.9)
Yes	60 (14.4)	23 (13.1)	37 (15.3)	13 (14.0)	47 (14.5)
Hypertension **					
No	179 (42.9)	67 (38.3)	112 (46.3)	31 (33.3)	148 (45.7)
Yes	235 (56.4)	107 (61.1)	128 (52.9)	61 (65.6)	174 (53.7)
Liver disease **					
No	403 (96.6)	172 (98.3)	231 (95.5)	91 (97.8)	312 (96.3)
Yes	12 (2.9)	3 (1.7)	9 (3.7)	2 (2.2)	10 (3.1)
Smoking Status **					
Current Smoker	13 (3.1)	6 (3.4)	7 (2.9)	5 (5.4)	8 (2.5)
Ex-smoker	47 (11.3)	21 (12.0)	26 (10.7)	11 (11.8)	36 (11.1)
Never smoker	233 (55.9)	98 (56.0)	135 (55.8)	40 (43.0)	193 (59.6)
Dataset	**Testing Cohort (Mount Sinai)**				
COVID-19 Severity		Severe		Critical	
	All Samples (n=569)	Cases (n=392)	Controls (n=177)	Cases (n=233)	Controls (n=336)
Age in years *	59.6 (19.4)	63.0 (17.1)	52.1 (22.1)	63.7 (16.7)	56.8 (20.6)
Female sex **	238 (41.8)	154 (39.3)	84 (47.5)	87 (37.3)	151 (44.9)

Table 4.1: Demographic characteristics of the participating cohorts. Notes are listed in Table 4.2

* Mean (SD)
** N (%). When the counts of each cell do not sum to the total sample size, this is due to missing data or the patient answering, “I do not know”.
Severe - Cases: individuals who tested positive for COVID-19 and died or required any type of respiratory support (including oxygen delivered by nasal prongs). Controls: individuals with COVID-19 who did not meet severe case criteria or individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative.
Critical - Cases: individuals who tested positive for COVID-19 and died or required respiratory support (intubation, continuous positive airway pressure, bilevel positive airway pressure, continuous external negative pressure, or high flow positive end expiratory pressure oxygen). Controls: individuals with COVID-19 who did not meet critical case criteria or individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative.
Sample Processing Time: time in hours between sample collection and sample freezing
CHUM: Samples recruited from the Centre Hospitalier de l’Université de Montréal
JGH: Samples recruited from the Jewish General Hospital

Table 4.2: Notes for Table 4.1

The clinical variables were collected for the BQC19 cohort only. Clinical variables included smoking status and six different comorbidities: diabetes, COPD, chronic kidney disease, congestive heart failure, hypertension, and liver disease. All seven variables were collected as categorical values with the six comorbidities having three options (0 No, 1 Yes, and -1 Don’t know) while smoking status contained 4 categories (0 Current Smoker, 1 Ex-smoker, 2 Never smoked, and -1 Don’t know).

4.1.3 Proteomic Measurement using the Somascan Platform

Blood samples from both the BQC19 and Mount Sinai cohorts were collected using acid citrate dextrose (ACD) tubes. Proteomic measurement was performed at Somalogic using the Somascan v4.0 platform. In the BQC19 cohort, a total 1,038 samples collected at up to five different time points from 503 individuals were sent to SomaLogic for proteomic profiling as previously described [4], while the Mount Sinai cohort contained 1,200 samples collected at multiple time points from 592 individuals that were sent to SomaLogic for proteomic profiling.

SomaLogic uses the Somascan proteomic platform which provides measurements on 4,701 unique human circulating proteins using 4,984 Slow Off-Rate Modified Aptamers (SOMAmer reagents) and quantifies protein levels in the form of relative fluorescence units (RFUs). Normalization and calibration steps were performed by SomaLogic to remove any systematic biases stemming from raw assays or samples. The normalization procedure involved three steps performed in a non-consecutive fashion: hybridization control normalization, intraplate median signal normalization, as well as plate scaling and calibration. More details on SomaLogic normalization can be found in their Technical Note [43].

4.1.4 SomaLogic Quality Control and Normalization

SomaLogic normalization is generally performed using the normal population reference derived from Ethylenediaminetetraacetic acid (EDTA) plasma samples. Due to plasma samples in the BQC19 cohort being collected using acid citrate dextrose (ACD) tubes, instead of EDTA tubes, the samples were normalized against the median of uncalibrated, unnormalized results, and this normalization reference was applied across all samples and plates. After normalization, samples are expected to lie in the acceptance range of [0.4-2.5] and samples outside this range are flagged. Normalization scaling values on the lower end of the acceptance range (closer to 0.4) reflect higher concentration of plasma proteins compared to values on the upper end of the acceptance range (2.5). Interestingly, we observed a higher proportion of flagged samples than usual compared to previous SOMAscan studies. A large starting blood volume in ACD tubes as well as discrepancies in the blood volume may explain the higher rates of flagging on the upper end of the normalization scaling acceptance range (low signal flags). Further, COVID-19 severity is associated with higher rates of flagging on the lower end of the acceptance range (high signal flags) which may in part be due to excessive amounts of proteins present in extremely ill patients undergoing an acute inflammatory response. (Note: samples collected from the CHUM hospital site of the BQC19 generally consisted of hospitalized and thus more severe COVID-19 patients).

SomaLogic data is provided in two forms: a normalized dataset and another dataset without sample normalization. For this project, we used the sample normalized dataset although further exploration evaluating alternative methods for normalization may be undertaken in the future.

4.1.5 Data Preprocessing

A per-sample normalization process involved using a scale factor for a set of SOMAmer reagents to compute against a reference value generated from the median of all calibrated, unnormalized samples, and then aggregating the results within a dilution. This was done because using a normal population reference generated from EDTA plasma tubes would have been inappropriate for normalization, since samples in this study were from ACD plasma tubes. Due to the nature of the samples that were collected from patients during acute infection, we did not apply the recommended scale [0.4-2.5] to remove samples. The raw dataset composed of 5,284 SOMAmer reagents was first processed with SomaLogic package SomaDataIO v3.1.0. We removed any SOMAmer reagents that represented non-human proteins or controls (NoneX, NonHuman, Spuriomer, HybControlElution, NonBiotin, NonCleavable) and retained 4,984 unique SOMAmer reagents for analysis.

4.1.6 Curation of Samples from the Longitudinal Dataset

To investigate our primary study question, we focused on samples collected during the acute infection stage. Samples from the acute infection stage were defined as samples collected from SARS-CoV-2 PCR positive patients within 14 days of symptom onset. When an individual provided multiple samples collected within 14 days of symptom onset, the sample collected at the earliest timepoint was retained for analyses. We chose to use samples close to symptom onset to reflect the proteome of acute COVID-19, rather than its recovery phase. Both the BQC19 and Mount Sinai samples adhered to these rules.

4.1.7 COVID-19 Outcome Definitions

We defined two sets of severity outcomes for COVID-19: **severe** COVID-19 and **critical** COVID-19. Positive SARS-CoV-2 results were confirmed by SARS-CoV-2 viral nucleic acid amplification tests (NAAT) from relevant biologic fluids.

Cases for severe COVID-19 were defined as individuals who tested positive for COVID-19 and died or required any type of respiratory support (including oxygen delivered by nasal prongs) at any timepoint. Controls for severe COVID-19 were defined as individuals who did not meet these severe case criteria; thus, controls were individuals with COVID-19 but did not meet severe case criteria or were individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative. This definition for controls is consistent with recent successful large-scale genetic studies [5].

Cases for critical COVID-19 were defined as individuals who tested positive for COVID-19 and died or required invasive respiratory support (intubation, continuous positive airway pressure, bilevel positive airway pressure, continuous external negative pressure, or high flow positive end expiratory pressure oxygen) at any timepoint. Controls for critical COVID-19 were individuals with COVID-19 but did not meet critical case criteria or were individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative.

4.1.8 Multivariable Logistic Regression

Multivariable logistic regression models were used to test the associations of either severe or critical COVID-19 on four covariates along with each SOMAmer reagent: age, sex, sample processing time, and hospital site. We used R package "glm" to perform 4,984 logistic regression models in the BQC19 cohort. We first applied a false discovery rate of

$P < 0.01$ (corrected P values were determined using the Benjamini-Hochberg procedure [67], p.adjust with method set to "BH" in R) to select a subset of proteins associated with severe or critical COVID-19 outcomes. Volcano plots measuring the uncorrected -log_{10} P values as a function of the effect size estimates of each SOMAmer reagent were generated using the bioinfokit version 2.0.4 package in Python 3.7.

4.1.9 Regularized Logistic Regression Models

We defined two model types differing in the covariates used to train the model. The first model type is a "baseline model" which was trained using age, sex, sample processing time, and hospital site. We then evaluated whether the addition of proteins would improve the identification of which patients developed adverse COVID-19 outcomes by adding 4,984 SOMAmer reagents to the baseline model. This second model type is a "protein model" which is trained using age, sex, sample processing time, hospital site, and 4,984 SOMAmer reagents. We used the baseline model in our analyses as a performance benchmark to compare the results of the protein model which we expected to perform better.

To predict the two COVID-19 severity outcomes defined above, we used LASSO regression [85] and elastic nets [87]. Specifically, for LASSO regression we used L_1 Regularized Logistic Regression (Sparse Logistic Regression) as implemented in the "LogisticRegression" module from Sci-kit learn version 0.24.1 [98], a machine learning library, with the penalty set to "L_1". The L_1 norm penalty adds a constraint to the effect estimates of the regression model by setting many variables to have a null effect or a coefficient of 0. This in turn allows a form of feature selection to occur and also prevents overfitting to the training dataset by forcing the model to be less complex. In addition, when multiple variables are correlated with one another such as in the case of highly correlated proteins, the penalty term from LASSO may select a single variable from the group, thus allowing a subset of uncorrelated proteins to be selected. It is important to note that although LASSO tends to select a single variable from a group of highly correlated variables, this property is not a certainty. LASSO may occasionally select more than one variable depending on the size of the dataset and the value of the penalty term. To train this model, the hyperparameter "λ", which controls the amount of L_1 regularization to add to the model, was first tuned through cross-validation as per equation 3.4 (details are described in the next section). A larger value of λ increases the amount of L_1 regularization and forces more of the variables to have a null effect. On the other hand, training the model on a smaller λ value will result in a model with more nonzero coefficients.

4.1.10 Using Elastic Net to Select Uncorrelated Proteins

For elastic net, we used elastic net regularized logistic regression (linear combination of L_1 and L_2 regularization). To implement this elastic net, we used the "LogisticRegression" module from sci-kit learn with penalty set to "elasticnet". The elastic net penalty combines both the L_1 norm and the L_2 norm penalties. The L_1 regularization term forces coefficients to be zero and in effect performs feature selection (as explained above). The L_2 regularization term does not force coefficients to have a null effect but instead reduces their magnitude, resulting in a model with a greater number of variables but many with small effect estimates. Therefore, the combination of both of these regularization terms gives rise to a more generalized form of the L_1 regularized logistic regression model used in the main analysis. For model training, the first step requires tuning two hyperparameters. In addition to the hyperparameter λ which controls the amount of L_1 regularization as previously described for LASSO logistic regression, the hyperparameter, "α", must also be tuned as per equation 3.6. In Sci-kit learn, this parameter is termed the "L_1 ratio" and ranges between 0 and 1 while controlling the amount of L_1 to L_2 regularization. For example, if the L_1 ratio is set to 1, this is in essence the L_1 regularized logistic regression model. Setting the L_1 ratio to the other extreme, 0, results in an L_2 regularized logistic regression model which is termed Ridge Logistic Regression. The tuning of these two hyperparameters was also performed through cross-validation. For the λ hyperparameter, we searched over the same range of values as that used in L_1 regularized logistic regression in the main analysis. For the L_1 ratio, we searched over 11 values: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0].

For the elastic net regularized logistic regression model, we used a tolerance parameter of 0.01 during training instead of 0.0001 used for L_1 regularized logistic regression to speed up tuning. For reproducibility, the random seed was set to 0 for cross-validation splits as well as weight initialization during model training. The training performance between L_1 and elastic net logistic regression was generally similar, so we decided to use the L_1 regularized logistic regression model to test on the external cohort due to it being faster to train and less complex than elastic net. We used the best hyperparameter λ, selected from cross-validation, to train the baseline and protein model using the entire BQC19 cohort. We pre-processed the training dataset by natural log transforming the protein levels due to protein levels varying considerably. The binary predictor variables sex and hospital site were dummy encoded while the continuous variables age and sample processing time were left as is. All protein variables were standardized to a mean of zero and unit variance.

4.1.11 Cross-Validation and Hyperparameter Tuning

Due to the relatively small size of our training dataset from BQC19, we used 10 repeats of five-fold cross-validation to tune the hyperparameter, λ, over 17 different λs (log_{10} values of λ from -2 to 2, incremented by 0.25). Each repeat of the five-fold cross-validation process involved splitting the dataset into five folds: training on four folds and validating the trained model on the final fold and performing the process five times to cover each validation fold. We used a stratified cross-validation approach, meaning that the train and validation folds maintained the same percentage of samples of each class (case/control) as the original data. This is important because of the unbalanced case/control samples for the critical COVID-19 outcome (93 cases / 324 controls). A standard five-fold cross-validation split may result in train and validation folds with varying proportions of cases and controls. Since classification algorithms tend to weigh each sample equally, the class that is overrepresented, such as the controls in the critical COVID-19 outcome, will receive more weight and thus may bias the results. The stratified cross-validation step was performed using the RepeatedStratifiedKFold function in Sci-kit learn. Due to the relatively small sample size of our training set (n=417), we performed 10 repeats of this cross-validation process to stabilize the results from training. Each repeat first shuffled the entire training set then split the data into five folds which created more variability in the data used for training.

During training on the four folds, we standardized only the protein levels using the population standard deviation (i.e., dividing by the number of samples n) and use this mean and standard deviation to standardize the protein levels in the validation fold prior to validating. This prevents information leakage which can occur if standardization of protein levels was performed on the entire dataset rather than just the training folds. Age and sample processing time were treated as continuous variables, whereas sex and hospital site were treated using dummy variables (sex [0: Female, 1: Male], and hospital site [0: CHUM, 1: JGH]).

We used the area under the curve (AUC) to determine the model performance during cross-validation (as described in section 3.5.3). To select the best value for the hyperparameter λ, we compared the average AUC score (computed from 50 validation fold results) for all λ values and selected the λ value corresponding to the highest average AUC. Youden's J statistic was calculated for each receiver operator characteristic (ROC) curve during training. This performance metric can be calculated by subtracting the false positive rate from the true positive rate for each data point on a ROC curve and taking the maximum value. The threshold which corresponds to this maximum Youden's J statistic is the threshold that maximizes the sum of the sensitivity and specificity for that particular ROC curve. We computed the threshold corresponding to the maximum Youden's J statistic for each of the 50 ROC curves and averaged the 50 thresholds to get a single threshold value. This averaged threshold value was computed for each of the baseline and protein models predicting severe and critical COVID-19 and used to produce two-by-two contingency tables and therefore sensitivity and specificity values in the Mount Sinai cohort during model testing.

4.1.12 Model Testing

We checked the generalizability of the baseline and protein models by testing them in an external, independent dataset from Mount Sinai. Protein measurements in the test dataset were first natural log transformed then standardized using the mean and standard deviation of the corresponding protein in the training set. Similarly, age was not standardized and kept in years. Since samples were only from a single hospital, the hospital site parameter was left as is and did not need to be dummy encoded. The variable sample processing time, however, was absent from the testing set. For this reason, we imputed the sample processing time variable in the test cohort using the mean value of the sample processing time variable in the BQC19 training cohort.

4.1.13 Protein Correlations

Spearman's Rank Order Correlation was used to determine the correlations between individual proteins in the BQC19 cohort. Heatmaps show magnitudes of correlation coefficients between values of -1 and 1. Correlation heatmaps showing collected clusters such as in Figures 4.8, 4.10, and 4.11 were generated using the ggcorrplot function in R with the parameter hc.order set to TRUE to perform hierarchical clustering. Moreover, we reduced the dimension of the correlation matrix of 4,984 SOMAmer reagents to a 2-dimensional space using uniform manifold approximation and projection (UMAP) from the umap-learn 0.5.1 package using default parameters. We annotated the SOMAmer reagents selected from the protein model that were associated with severe COVID-19 and critical COVID-19 as well as the proteins that overlapped between the two outcomes.

4.1.14 Pathway Enrichment Analyses

Pathway enrichment analysis provides insights into the mechanisms behind genes. By providing a gene list to this method, it is able to determine biological pathways that may be more significantly enriched than by chance alone, thus enabling discovery into the mechanisms of groups of genes generated from commonly performed omics experiments. In this project, we used the web-based tool g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) to investigate the possible pathways of the selected proteins as good predictors for both critical and severe COVID-19 identified by LASSO. The g:SCS algorithm was used to estimate the threshold for enrichment against all annotated genes. It is used to compute P values adjusted for multiple testing after performing pathway enrichment analysis. To produce the results in Supplementary Figure 4.13, we selected pathways and interaction databases including Gene Ontology [99], KEGG [100], Reactome [101], TRANSFAC [102], miRTarBase [103], Human Protein Atlas (proteinatlas.org) [104], and CORUM [105].

4.1.15 Sensitivity Analyses

We tested the effect of six established clinical risk factors which included: diabetes, COPD, chronic kidney disease, congestive heart failure, hypertension, and liver disease in the BQC19 cohort to determine whether addition of comorbidities could improve prediction of COVID-19 severity outcomes. We added these six additional covariates, with characteristics shown in Table 4.1, to the baseline and protein models to perform LASSO regression analysis. A total of 417 samples from the BQC19 cohort were used for training.

We performed a second sensitivity analysis by adding smoking status along with these six established clinical variables to the baseline and protein models for LASSO regression analyses. Therefore, the baseline model contained covariates age, sex, sample processing time, hospital site, and seven clinical variables while the protein model contained all the baseline variables along with 4,984 SOMAmer reagents. Since smoking status was not available from the CHUM hospital site, this sensitivity analysis only involved 312 samples from the BQC19 cohort that were collected at the JGH site.

Due to missing data (displayed in Table 4.3), we imputed the values of samples: we first converted all six comorbidity features to binary values. Any value other than a "Yes" was converted to a "No" which may include missing values being converted to a "No". For smoking status, we grouped all values into three categories: 0 - Current Smoker, 1 - Ex-smoker, and anything else (including missing values and -1) was set as 2 - Never smoked. Smoking status was dummy encoded and had one of the encoded variables dropped to prevent collinearity. For both sensitivity analyses, training of the L_1 regularized logistic regression models used 10 repeats of stratified five-fold cross-validation as in the primary analysis.

4.2 Experimental Results

4.2.1 Cohorts

The demographic and clinical characteristics of the participants in the training and testing datasets are shown in Table 4.1. In the BQC19 cohort, the mean age across all samples was 65.3 years (SD = 18.4 years), and 52% of the cohort were men. In the Mount Sinai cohort, the mean age was 59.6 years (SD: 19.4 years), and 58.2% of the cohort were men.

For the definition of adverse COVID-19 outcomes, we focused on two levels of severity as described in section 4.1.7. The overall study design is shown in Figure 4.1, which outlines the training and testing stages of the study. In the BQC19 training cohort 175 individuals were classified as severe cases and 242 individuals were controls. The controls for severe COVID-19 were comprised of 138 SARS-CoV-2 positive individuals not meeting case definition and 104 SARS-CoV-2 negative individuals. In the case of critical disease, 93 individuals out of 313 COVID-19 positive patients were classified as critical cases and 324 individuals were controls. The controls for critical COVID-19 cases were 220 SARS-CoV-2 positive individuals not meeting case definition and 104 participants

Dataset	Training Cohort (BQC19)				
COVID-19 Severity		**Severe**		**Critical**	
	All Samples (n=417)	Cases (n=175)	Controls (n=242)	Cases (n=93)	Controls (n=324)
Diabetes **					
No	288 (69.1)	100 (57.1)	188 (77.7)	54 (58.1)	234 (72.2)
Yes	127 (30.5)	75 (42.9)	52 (21.5)	39 (41.9)	88 (27.2)
Patient answering “I do not know”	2 (0.5)	0 (0.0)	2 (0.8)	0 (0.0)	2 (0.6)
Chronic Obstructive Pulmonary Disease **					
No	358 (85.9)	145 (82.9)	213 (88.0)	77 (82.8)	281 (86.7)
Yes	57 (13.7)	30 (17.1)	27 (11.2)	16 (17.2)	41 (12.7)
Missing	2 (0.5)	0 (0.0)	2 (0.8)	0 (0.0)	2 (0.6)
Chronic Kidney Disease **					
No	363 (87.1)	151 (86.3)	212 (87.6)	77 (82.8)	286 (88.3)
Yes	52 (12.5)	24 (13.7)	28 (11.6)	16 (17.2)	36 (11.1)
Missing	2 (0.5)	0 (0.0)	2 (0.8)	0 (0.0)	2 (0.6)
Congestive Heart Failure **					
No	355 (85.1)	152 (86.9)	203 (83.9)	80 (86.0)	275 (84.9)
Yes	60 (14.4)	23 (13.1)	37 (15.3)	13 (14.0)	47 (14.5)
Missing	2 (0.5)	0 (0.0)	2 (0.8)	0 (0.0)	2 (0.6)
Hypertension **					
No	179 (42.9)	67 (38.3)	112 (46.3)	31 (33.3)	148 (45.7)
Yes	235 (56.4)	107 (61.1)	128 (52.9)	61 (65.6)	174 (53.7)
Missing	3 (0.7)	1 (0.6)	2 (0.8)	1 (1.1)	2 (0.6)
Liver disease **					
No	403 (96.6)	172 (98.3)	231 (95.5)	91 (97.8)	312 (96.3)
Yes	12 (2.9)	3 (1.7)	9 (3.7)	2 (2.2)	10 (3.1)
Missing	2 (0.5)	0 (0.0)	2 (0.8)	0 (0.0)	2 (0.6)
Smoking Status **					
Current Smoker	13 (3.1)	6 (3.4)	7 (2.9)	5 (5.4)	8 (2.5)
Ex-smoker	47 (11.3)	21 (12.0)	26 (10.7)	11 (11.8)	36 (11.1)
Never smoker	233 (55.9)	98 (56.0)	135 (55.8)	40 (43.0)	193 (59.6)
Patient answering “I do not know”	124 (29.7)	50 (28.6)	74 (30.6)	37 (39.8)	87 (26.9)

Table 4.3: Rate of missingness for seven clinical variables in the BQC19 training cohort.

who were SARS-CoV-2 negative. In the Mount Sinai testing cohort, 392 individuals were classified as severe cases and 177 individuals were controls while for critical disease 233 individuals were cases and 336 were controls. Generally, severe, or critical COVID-19 cases were older than controls in both the training dataset and the testing dataset. Males were also more likely to have severe or critical COVID-19 as compared to females (Table 4.1). The age and sex distribution of the participants stratified by case/control status for the two COVID-19 severity outcomes are shown in Figure 4.2. The distributions suggest that males who develop severe or critical COVID-19 are generally younger than females.

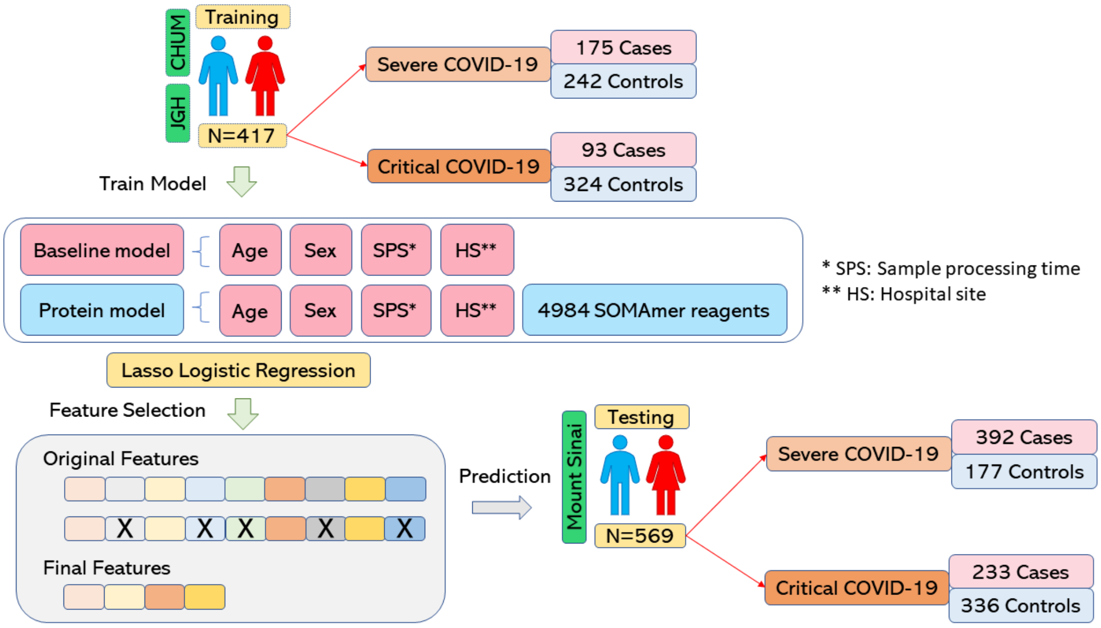
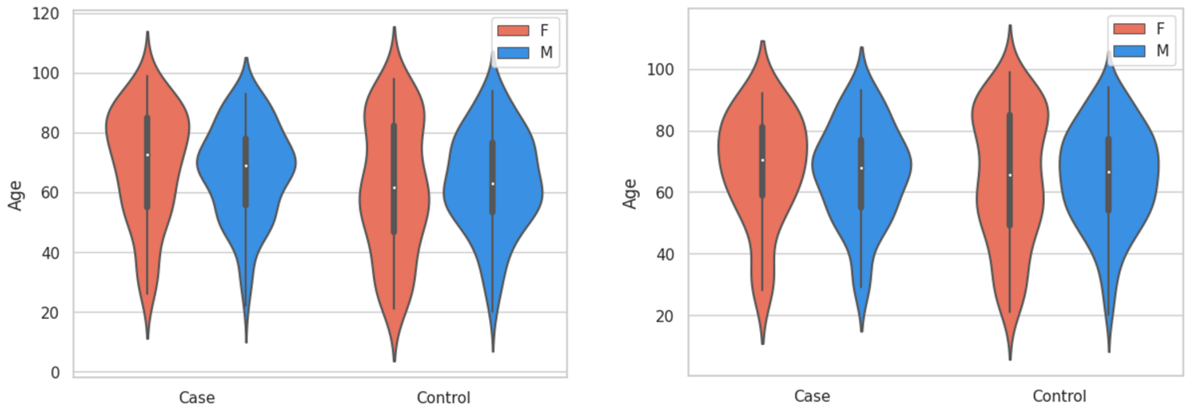


Figure 4.1: Overall study design. Schematic of training and testing stages of this study. Severe COVID-19 is defined as death or use of any form of oxygen supplementation. Critical COVID-19 is defined as death or severe respiratory failure (non-invasive ventilation, high flow oxygen therapy, intubation, or extracorporeal membrane oxygenation).

4.2.2 Association of Protein Abundance with COVID-19 Outcomes

In order to directly assess if any of the measured proteins were associated with COVID-19 severity, we used multivariable logistic regression to test the association of each of the 4,984 SOMAmer reagents with the two COVID-19 outcomes while adjusting for age,



(a) Severe COVID-19

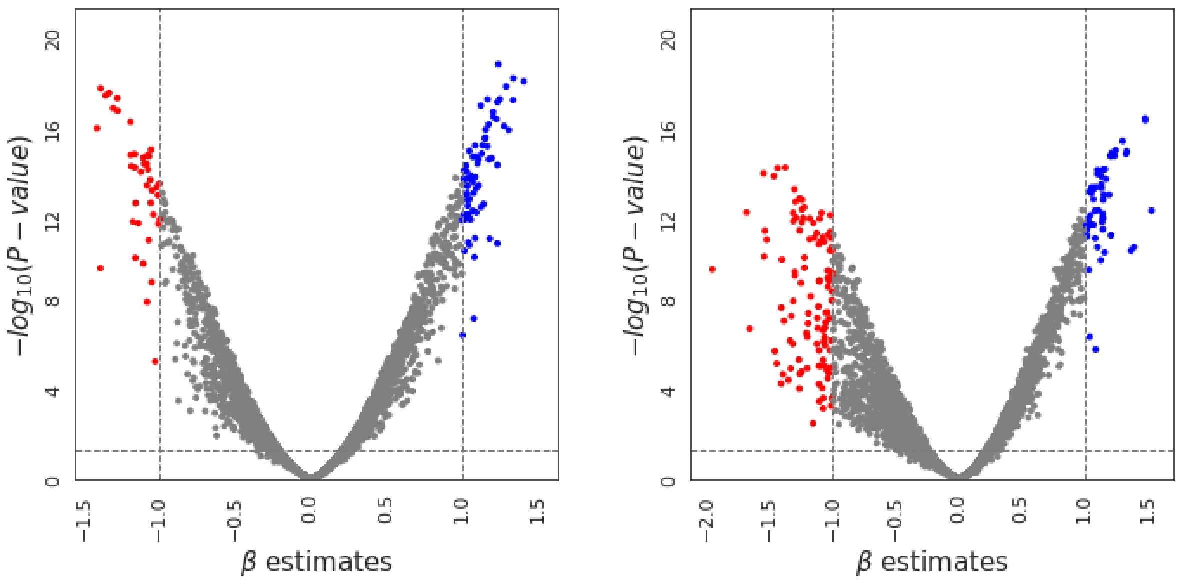
(b) Critical COVID-19

Figure 4.2: Violin plots showing the age and sex distribution of the BQC19 training cohort for severe and critical COVID-19. a, Severe COVID-19. b, Critical COVID-19. Coloring indicates the sex of the patient as shown in the legend (F: Female, M: Male).

sex, sample processing time, and hospital site in the BQC19 cohort. These variables were chosen because they are readily available in the course of clinical care, representing the minimum set of variables to predict outcomes. Logistic regression identified 1,531 SOMAmer reagents to be associated with severe COVID-19 (Appendix A - Supplementary Table 1) and 1,592 SOMAmer reagents (Appendix A - Supplementary Table 2) to be associated with critical COVID-19 when using a Benjamini-Hochberg corrected P value of 0.01 (Figure 4.3). Within these two groups of associated proteins, 1,264 proteins are associated with both severe and critical COVID-19 outcomes, 328 proteins are only associated with critical COVID-19, and 267 proteins are only associated with severe COVID-19 for a total of 1,859 uniquely identified proteins.

4.2.3 Model Selection and Performance Using LASSO

To train both the baseline and the protein models (described in section 4.1.9), we performed 10 repeats of stratified 5-fold cross-validation using LASSO logistic regression in the BQC19 cohort on both the severe and critical outcomes. We tuned the penalty parameter "λ" across each of the 50 cross-validations and selected the λ value corresponding to



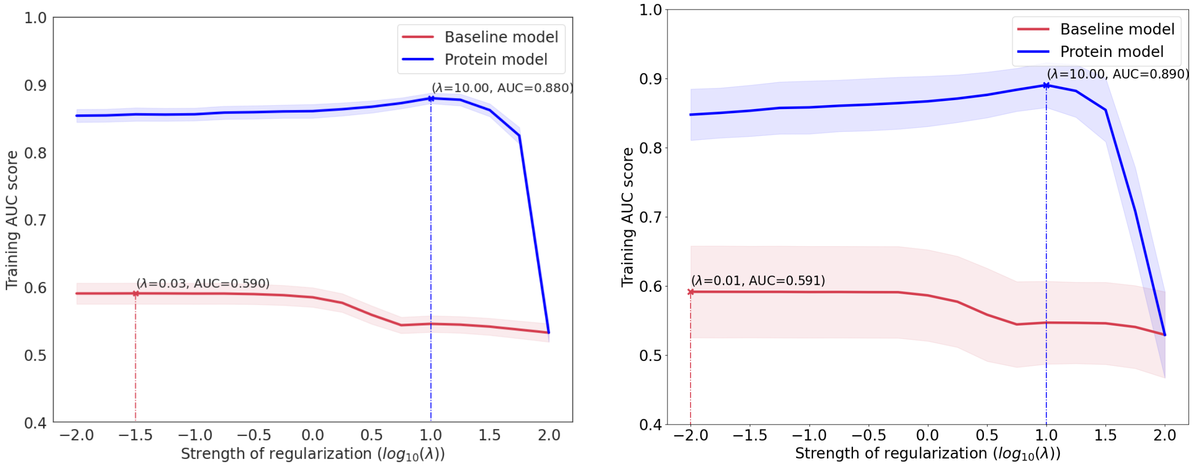
(a) Severe COVID-19

(b) Critical COVID-19

Figure 4.3: Volcano plots. a, Severe COVID-19. b, Critical COVID-19. Each data point represents the association of a single SOMAmer reagent from multivariable logistic with these COVID-19 outcomes adjusted for age, sex, sample processing time, and hospital site. Vertical gridlines are set at β effect estimates of -1.0 and 1.0 while the horizontal gridline represents the P value with a significant threshold set at P = 0.05. The β effect estimate shows the change in the log odds of having severe or critical COVID-19 disease associated with a single standard deviation change in the SOMAmer reagent of interest. Statistically significant associations between SOMAmer reagents and COVID-19 severity with large β estimate changes in the negative direction and positive direction are shown in red and blue, respectively.

the model with the highest area under the receiver operator characteristic curve (AUC), which was averaged over the 50 cross-validation results. Results from the λ parameter search are shown in Figure 4.4.

For the best performing model predicting the severe COVID-19 outcome, we selected a $log_{10}\ \lambda$ value of -1.5 which generated an average training AUC of 59% for the baseline model. We next selected a $log_{10}\ \lambda$ value of 1.0, which generated an average AUC of 88% for the protein model. For the best performing model predicting the critical COVID-19 outcome, we selected $log_{10}\ \lambda$ values of -2.0 and 1.0 corresponding to average cross-validation



(a) Severe COVID-19 outcome

(b) Critical COVID-19 outcome

Figure 4.4: AUC score as a function of regularization strength during L_1 regularized logistic regression training to determine the best λ hyperparameter value. Results presented show model training on the BQC19 cohort. a, Severe COVID-19 outcome. b, Critical COVID-19 outcome. The baseline model (red) was fitted with covariates age, sex, sample processing time, and hospital site. The protein model (blue) was fitted with age, sex, sample processing time, hospital site, and 4,984 SOMAmer reagents. Each data point represents an AUC score averaged over 50 validation folds. Shaded areas represent 95% confidence intervals above and below the mean AUC.

training AUC scores of 59% and 89% for the baseline and protein model, respectively (Figure 4.5). We then used these chosen λ hyperparameters to build baseline and protein models for severe and critical COVID-19 using the entire BQC19 cohort and evaluated their performance in the independent external test cohort from Mount Sinai.

[1] Footnote for Figure 4.5. Standard deviation was calculated assuming the *population standard deviation* σ using pandas.DataFrame.std(ddof=0) which is consistent with numpy.std. To calculate confidence intervals for the population mean μ, we use the sample mean $\bar{x}$ as follows

$$\mu = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

where, σ is the population standard deviation, n is the sample size, and for 95% confidence intervals, $z = 1.96$. Note that $\frac{\sigma}{\sqrt{n}}$ is the standard error.

Here, we calculate 95% CI for the training AUC curves where $n = 50$ since each data point is averaged over 50 validation set AUC score.

The 95% CI says that only 5% of the time, the mean will not lie in the range

$$\mu = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

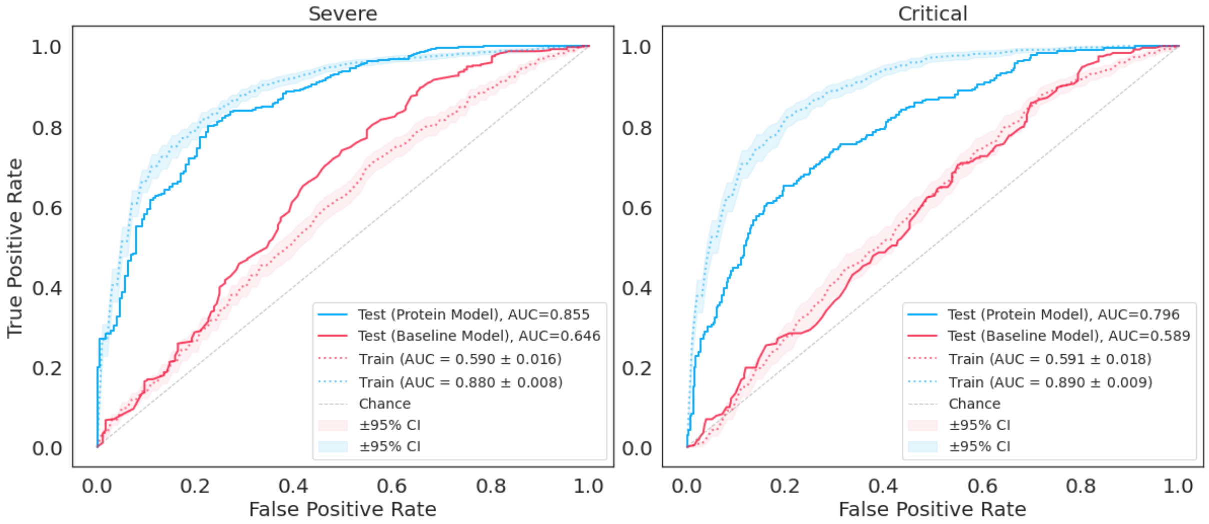


Figure 4.5: AUC score results. (Left), L_1 regularized logistic regression training and testing results for severe COVID-19 and (Right), critical COVID-19. Blue and red are used to represent the protein model and baseline model, respectively, while solid and dotted lines represent the testing and training performance, respectively. Shaded areas denote the 95% confidence intervals for the training cohort.[1]

When testing the prediction of severe COVID-19 in the independent Mount Sinai cohort, AUC performance of the baseline model improved from 59% in the BQC19 training cohort to 65% in the Mount Sinai testing cohort. The AUC of the protein model decreased slightly between training and testing (88% vs. 86%).

The AUC of the protein model for predicting critical COVID-19 also decreased from a training score of 89% to 80% in the test set. In contrast, prediction of the critical COVID-19 outcome using the baseline model was consistent between training and test performance (AUC: 59%). The stability of these AUC estimates in the test cohort suggested that both protein models were robust.

The classification performance of the baseline and protein models in the Mount Sinai cohort is shown as two-by-two contingency tables in Figure 4.6. The baseline and protein models used thresholds of 0.417 and 0.486 to predict severe COVID-19, respectively. These thresholds were selected by computing Youden's J statistic during training and determine the threshold that maximized the sum of the sensitivity and specificity scores during training. The thresholds selected were roughly consistent with the case to control ratio in the BQC19 cohort used for training (175 cases, 242 controls). The protein model achieved a sensitivity of 73.2% compared to 61.0% for the baseline model, and a specificity of 79.7% compared to 60.5% for the baseline model, when predicting the severe COVID-19 outcome (Figure 4.6).

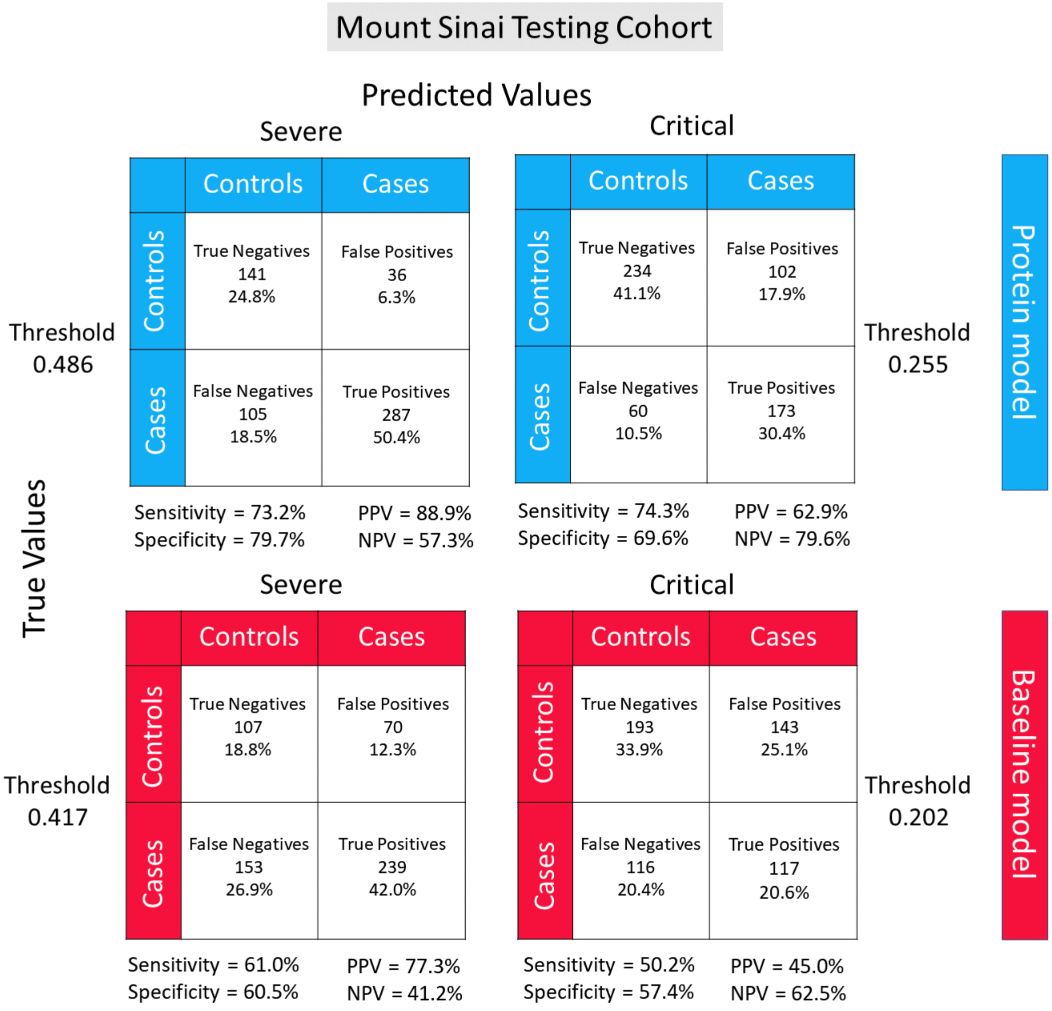
When predicting critical COVID-19 using the baseline and protein models, thresholds of 0.202 and 0.255 were used to predict cases, respectively, using the same method. The low threshold for predicting critical COVID-19 cases is consistent with the case to control ratio in training which was 93 to 324 samples. The baseline model achieved a sensitivity / specificity score of 50.2% / 57.4% while the protein model achieved 74.3% / 69.6%, suggesting that the protein model trained to predict critical COVID-19 had decent power to classify true positives and true negatives (Figure 4.6).

Furthermore, both the baseline and protein models demonstrated higher positive predictive values than negative predictive values when predicting the severe COVID-19 outcome, compared to the critical outcome. In contrast, these models produced higher negative predictive values than positive predictive values when predicting critical COVID-19.

Overall, these results suggest that the protein models predicting severe and critical COVID-19 both perform reasonably well in terms of the trade-off between sensitivity and specificity. The protein model is sensitive (73.2%) at identifying severe COVID-19 cases and similarly sensitive (74.3%) at identifying critical COVID-19 cases. Further, the positive predictive value for severe COVID-19 was high at 88.9%, while the negative predictive value was 57.3% (Figure 4.6). These results suggest that a protein model could predict severe COVID-19 with relatively high confidence.

4.2.4 Training and Performance of Models using Elastic Net

The cross-validation results for elastic net are shown in (Figure 4.7). The best hyperparameters produced training AUCs of 87.3% and 88.5% for the protein model when predicting



Mount Sinai Testing Cohort

Predicted Values

True Values

Protein model

Severe — Threshold 0.486

	Controls	Cases
Controls	True Negatives 141 24.8%	False Positives 36 6.3%
Cases	False Negatives 105 18.5%	True Positives 287 50.4%

Sensitivity = 73.2% PPV = 88.9%
Specificity = 79.7% NPV = 57.3%

Critical — Threshold 0.255

	Controls	Cases
Controls	True Negatives 234 41.1%	False Positives 102 17.9%
Cases	False Negatives 60 10.5%	True Positives 173 30.4%

Sensitivity = 74.3% PPV = 62.9%
Specificity = 69.6% NPV = 79.6%

Baseline model

Severe — Threshold 0.417

	Controls	Cases
Controls	True Negatives 107 18.8%	False Positives 70 12.3%
Cases	False Negatives 153 26.9%	True Positives 239 42.0%

Sensitivity = 61.0% PPV = 77.3%
Specificity = 60.5% NPV = 41.2%

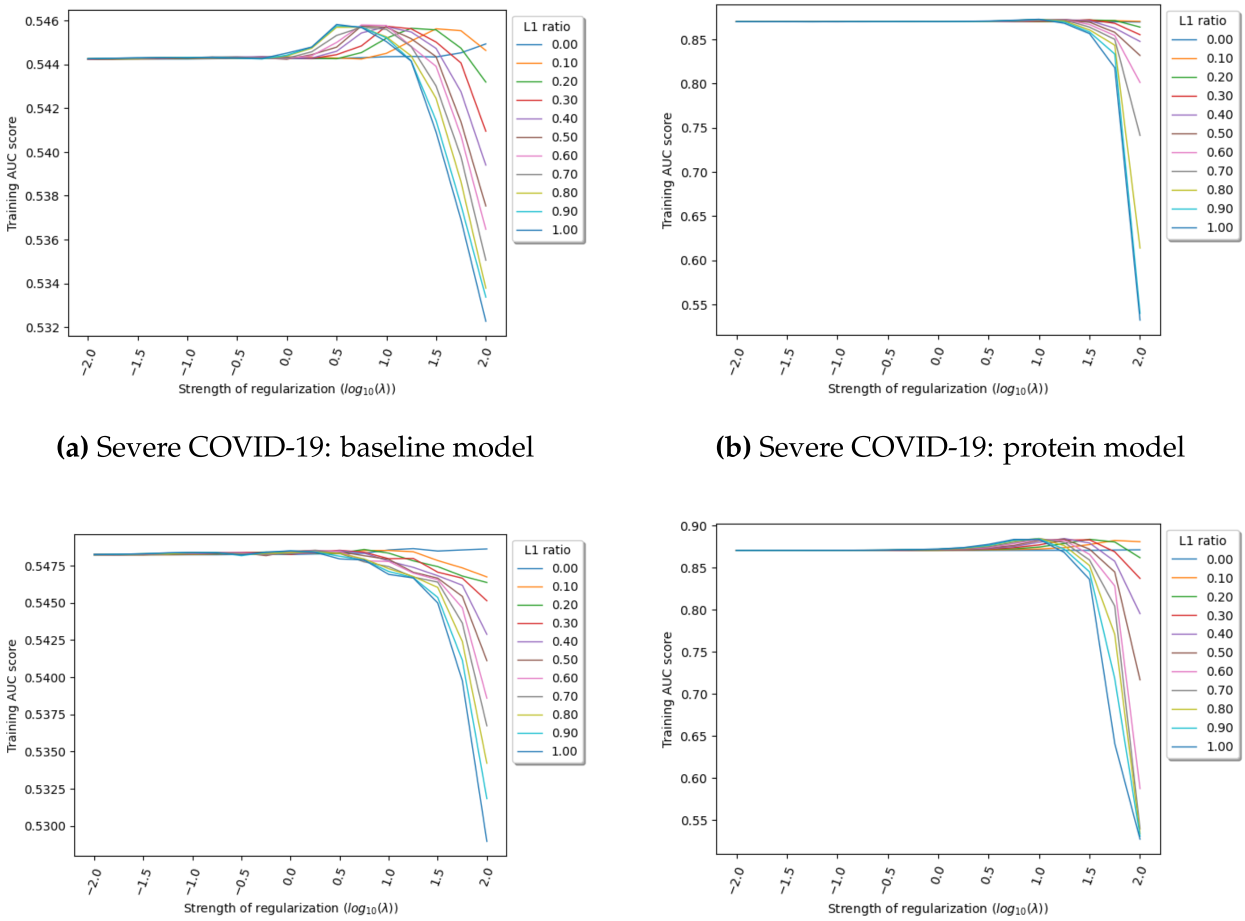
Critical — Threshold 0.202

	Controls	Cases
Controls	True Negatives 193 33.9%	False Positives 143 25.1%
Cases	False Negatives 116 20.4%	True Positives 117 20.6%

Sensitivity = 50.2% PPV = 45.0%
Specificity = 57.4% NPV = 62.5%

Figure 4.6: Two-by-two contingency table results from the test set are shown for predicting severe (top left, bottom left) and critical COVID-19 (top right, bottom right) using the protein model (blue) and baseline model (red). The threshold for predicting cases was determined during training using Youden's J statistic which selects a threshold that maximizes the sum of the sensitivity and specificity score. PPV = positive predictive value, NPV = negative predictive value.

severe COVID-19 and critical COVID-19, respectively (data not shown). For the baseline model, training AUC results for predicting severe COVID-19 and critical COVID-19 were 54.6% and 54.9%, respectively (data not shown). The training performance results for elastic net models were similar to LASSO models with the only improvement demonstrated by the protein model when predicting severe COVID-19 (LASSO training AUC



(a) Severe COVID-19: baseline model

(b) Severe COVID-19: protein model

(c) Critical COVID-19: baseline model

(d) Critical COVID-19: protein model

Figure 4.7: Elastic net logistic regression hyperparameter search results. AUC score as a function of regularization strength during elastic net regularized logistic regression training on severe COVID-19 for the a, baseline model and b, protein model. c, AUC score as a function of regularization strength during elastic net regularized logistic regression training on critical COVID-19 for the c, baseline model and d, protein model.

88.0% vs. elastic net training AUC 87.3%). Due to small differences between LASSO and elastic net performance, we did not pursue testing of the elastic net model in the external Mount Sinai cohort since it is more complex and requires more computational time to train.

4.2.5 Proteins Prioritized by LASSO to Predict COVID-19 Severity Outcomes

To predict the severe COVID-19 outcome, the best performing protein model selected 92 proteins along with age and sample processing time (Figure 4.9, Appendix A - Supplementary Table 3). Assessing the correlation of all 92 proteins, we found that, as expected, most of the proteins did not correlate with each other (mean absolute Spearman's ρ = 0.17) (Figure 4.8). Of 8,464 total correlations (92 × 92), 8,372 correlations (98.9%) had Spearman's absolute $\rho < 0.8$.

Next, when predicting the critical COVID-19 outcome, the best performing protein model retained age, sample processing time, and 67 proteins (Figure 4.9, Appendix A - Supplementary Table 4). The absolute effect estimates of these proteins were generally larger than the severe COVID-19 model proteins (mean: 0.081 vs. 0.077). As expected, the 67 selected proteins also showed low levels of correlation (mean absolute Spearman's ρ = 0.15) (Figure 4.10). Of 4,489 total correlations (67 × 67), 4,422 correlations (98.5%) had Spearman's absolute $\rho < 0.8$. The correlation between the 92 and 67 proteins selected to best predict severe and critical COVID-19 outcomes is shown in Figure 4.9. Out of 6,164 total correlations (92 × 67), 6,150 correlations (99.8%) had Spearman's absolute $\rho < 0.8$. In general, proteins selected for predicting severe versus critical COVID-19 were not highly correlated (mean absolute Spearman's ρ = 0.15). A hierarchically clustered heatmap after removal of the 14 common proteins in severe and critical COVID-19 showed that the proteins selected in predicting both outcomes were also generally uncorrelated with one another where 99.2% of the correlations had Spearman's absolute $\rho < 0.8$ (Figure 4.11).

The percent of selected proteins for the prediction of severe and critical COVID-19 that were cytokines or chemokines was only 5.4% and 4.5%, respectively. Cytokine IFNA7, as well as chemokines CXCL13, CXCL10, CCL7, and CCL8 were present in the proteins selected for predicting severe COVID-19. Three chemokines, CXCL13, CXCL10, and CCL7 were selected for predicting critical COVID-19. Importantly, among the 14 overlapping proteins, those three chemokines, CXCL13, CXCL10, and CCL7 were selected for predicting both severe and critical COVID-19.

In addition, we clustered the 4,984 SOMAmer reagents using uniform manifold approximation and projection (UMAP) to reduce the feature space to a 2-dimensional space and highlighted the position of the model-selected proteins predictive of either severe or critical COVID-19. Our results suggested that LASSO selected proteins were sparsely distributed across the clusters (Figure 4.12). This provides further evidence that: 1) few of the selected proteins are closely clustered with one another in UMAP space and 2) the proteins selected from the severe protein model and critical protein model were also quite distant from each other in UMAP space. Finally, we performed pathway analysis on common proteins with non-zero effect estimates that were included in the best predictive models of severe and critical COVID-19. Since LASSO is designed to pick one protein from a group of correlated proteins regardless of their biological relevance, we also included proteins highly correlated (Spearman's absolute $\rho > 0.75$) with the proteins selected by LASSO for the enrichment analysis. As a result, 171 proteins were included in the severe group (92 LASSO selected proteins and 79 correlated proteins, Appendix A - Supplementary Table 5); and 96 proteins were included in the critical group (67 LASSO selected proteins and 29 correlated proteins, Appendix A - Supplementary Table 6). Among which, 32 proteins were common between the severe and critical groups.

We found that these 32 proteins were enriched in 35 pathways (g:SCS adjusted P value < 0.05), among which 15 were directly related to immune responses (Figure 4.13, Appendix A - Supplementary Table 7). Prominent pathways included viral protein interaction with cytokines, cytokine and cytokine receptors (IL22RA1, TNFRSF10B, CCL7, CXCL10, CXCL13, adjusted P=0.0008) and cytokine-cytokine receptor interactions (CD4, IL22RA1, TNFRSF10B, CCL7, CXCL10, CXCL13, IFNA7, adjusted P= 0.002).

Interestingly, more than half of enriched pathways were not directly related to immune response (e.g., signaling receptor binding, cell activation) as shown in figure 4.13. This suggests that other non-immune response pathways influence COVID-19 severity.

In addition, some of these pathways included protein phosphorylation (CTSG, PRKCZ, PECAM1, CD4, CLEC7A, NPPA, TNFRSF10B, PRDX4, EPHA4, TNXB, CXCL10, IFNA7, CDH5) and glycosaminoglycan binding (CTSG, CCL7, TNXB, CXCL10, CXCL13), suggesting potential avenues to explore for drug development.

4.2.6 Using Clinical Risk Factors to Predict COVID-19 Outcomes

In order to contrast the prediction capabilities of protein levels with established clinical risk factors, we performed two sensitivity analyses with results shown in Table 4.4. In the first analysis, we added six clinical risk factors to the baseline and protein models described in the main analyses. These clinical risk factors were diabetes, chronic obstructive pulmonary disease (COPD), chronic kidney disease, congestive heart failure, hypertension, and liver disease. The prevalence of these risk factors is shown in Table 4.1.

Addition of these six clinical features to the baseline model improved the training AUC to 64% (from 59%) when predicting severe COVID-19 and to 61% (from 59%) when predicting critical COVID-19. However, adding these clinical risk factors to the protein model resulted in no change in the training AUC performance when predicting severe COVID-19 (AUC = 88% vs. 88%) and critical COVID-19 (AUC = 89% vs. 89%) (Table 4.4, Figures 4.14 and 4.15). 95 and 69 features with non-zero β coefficient effect estimates were selected for the protein models predicting severe and critical COVID-19, respectively, in this sensitivity analysis (Figures 4.14 and 4.15). Comparing proteins selected by the protein model in this sensitivity analysis, only one protein, KIT, was added to the 94 features selected in the main analysis. For critical COVID-19, the 69 features selected remained the same.

For the second sensitivity analysis, we augmented the first sensitivity analysis with an extra covariate for smoking status. Due to missing smoking information from the CHUM hospital site in the BQC19 cohort, only 312 samples were used in model training for the second sensitivity analysis. The results suggested that the addition of smoking and 6 clinical risk factors into the original baseline model composed of age, sex, sample processing time, and hospital site also slightly improved the training performance when predicting the severe COVID-19 outcome (AUC = 66% vs. 59%) and the critical COVID-19 outcome (AUC = 61% vs. 59%) (Table 4.4, Figures 4.16 and 4.17). When adding smoking and these 6 clinical risk factors to the protein model, we found that training performance actually decreased for the severe COVID-19 outcome (AUC = 85% vs. 88%) and critical COVID-19 outcome (AUC = 85% vs. 89%). The non-zero β coefficients (i.e., the effect sizes) of the proteins for severe and critical COVID-19 outcomes are shown in Figures 4.16 and 4.17 with a total of 79 and 51 features being selected, respectively. Comparing the 79 features selected by the protein model in this sensitivity analysis to the original 94 features selected previously when predicting severe COVID-19, we observed that only 48 features overlapped. Similarly, the 51 features selected by the protein model in this sensitivity analysis only had 28 features overlapping with the 69 features selected previously. The observed decrease in AUC and fewer number of overlapping proteins when comparing main analyses and sensitivity analyses may be due to the reduction in sample size used for training.

The results from these sensitivity analyses suggest that the protein measurements are likely able to act as partial proxies of the tested clinical risk factors. The addition of the clinical risk factors that we assessed may improve the predictive performance for both COVID-19 severity outcomes when only demographic and sample processing parameters are available. However, when protein measurements are available, adding these extra clinical risk factors may add little for improving predictions.

4.3 Discussion

In this section, we summarize and describe the implications of the work including limitations and novel insights in the context of the current literature on published COVID-19 proteomics studies.

	Main Analysis (n=417)		Main Analysis plus 6 clinical risk factors (n=417)		Main Analysis plus 7 clinical risk factors (n=312)	
	Baseline model	Protein model	Baseline model + 6 CRFs	Protein model + 6 CRFs	Baseline model + 7 CRFs	Protein model + 7 CRFs
Severe COVID-19	0.59	0.88	0.64	0.88	0.66	0.85
Critical COVID-19	0.59	0.89	0.61	0.89	0.61	0.85

Table 4.4: Training AUC comparison.
Main Analysis:
Baseline model: age, sex, sample processing time, and hospital site.
Protein model: baseline model and 4,984 SOMAmer reagents.

Main Analysis plus 6 clinical risk factors (CRFs):
Baseline model + 6 CRFs: age, sex, sample processing time, hospital site, and 6 comorbidities.
Protein model + 6 CRFs: baseline model + 6 CRFs and 4,984 SOMAmer reagents.

Main Analysis plus 7 clinical risk factors:
Baseline model + 7 CRFs: age, sex, sample processing time, hospital site, 6 comorbidities and smoking status.
Protein model + 7 CRFs: baseline model + 7 CRFs and 4,984 SOMAmer reagents.

4.3.1 Prediction of COVID-19 Outcomes

A major clinical challenge within the pandemic has been the triaging of patients to identify those most likely to require admission for hospitalization [106]. A common reason for hospitalization is the need for oxygen support. Currently, treating physicians are required to assess the need for admission using models with poor predictive performance. A model generated in China early in the pandemic to predict COVID-19 severity requires a medical history, chest X-ray, and extensive blood testing [107]. Further, the 4C Mortality Score was able to predict in-hospital mortality, but achieved an AUC of only 77% [108].

4.3.2 Prediction of COVID-19 Outcomes with Circulating Proteins

In this large-scale study testing the association of 4,701 circulating proteins with severe and critical COVID-19 outcomes, we found that a subset of these proteins were strong predictors of COVID-19 severity. Specifically, developing a model in 417 individuals and testing its performance in 569 separate samples from an independent external cohort, we demonstrated that a proteomic model was able to predict severe COVID-19, defined as requiring the use of oxygen, with an AUC of 86% and a positive predictive value of 89%. The addition of several commonly used clinical risk factors for COVID-19 severity did not improve the performance of this model. The identified proteins were strongly enriched for cytokine signaling and immune pathways, but also highlighted non-immune pathways. Taken together, these findings demonstrate that circulating protein abundances are able to predict COVID-19 outcomes with reasonable accuracy.

4.3.3 Inclusion of Secondary Cohort

By including an independent cohort in this study, we implemented best practices for model development and validation [46]. An important aspect of any prediction model is the testing of the model in a cohort separate from the training cohort. Therefore, a strength of this study was that our samples were recruited from three separate hospitals, across two separate health care systems in two different countries. In this study, we used the same proteomic measurement procedure to both train and test the models. This increases the probability that the results presented are generalizable and not overfitted to the training data [47]. Indeed, for the severe COVID-19 outcome, there was little change in the AUC when comparing the training and test cohorts (88% vs. 86%).

4.3.4 Circulating Cytokines and Chemokines

Further, most studies that have tested the association between protein levels and COVID-19 outcomes have focused on circulating cytokines and chemokines [16, 48–54]. While this is a reasonable approach given the nature of the disease, we are unaware of any other studies that have tested the association of 4,701 circulating proteins with COVID-19 outcomes. A recent study assessing thousands of proteins and their associations with COVID-19 outcomes achieved an AUC of 85% with an elastic-net logistic regression model, but this was not tested in an independent cohort [33]. This model included covariates such as age, gender, ethnicity, heart disease

Interestingly only 5 of the 14 proteins selected in the final model of both severe and critical COVID-19 outcome were cytokines or chemokines. There were also proteins selected that were not specific to immune pathway proteins, such as glycosaminoglycan binding—a favourable set of targets for drug development. This suggests that many of the biological pathways that influence severity of COVID-19 outcomes may act distinctly from known cytokine and chemokine proteins.

4.3.5 Limitations

This study has important limitations. While the model was tested in a separate cohort, and generalized well, it should be tested in additional cohorts, especially in cohorts of diverse ancestry. Due to sample size limitations, separate prediction models for each ancestry was not performed. The control population included individuals who were SARS-CoV-2 positive and had mild disease, in addition to individuals who were suspected to have COVID-19 but were SARS-CoV-2 negative. This means that the developed models provide insight into prediction of individuals who develop severe COVID-19 compared to mild COVID-19 and other acute diseases having symptoms consistent with COVID-19. Such control definitions reduce the potential for collider bias, but do not allow direct prediction of COVID-19 outcomes amongst only COVID-19 patients [109]. Another limitation of this study is that the identified proteins may only be surrogates of the causal proteins, since many are highly correlated. Thus, it is difficult to make causal inferences given these correlations. In addition, approximately half of the patients enrolled in the BQC19 have developed severe or critical COVID-19 after their baseline proteomic measurement which were used as predictors in this study. This suggests circulating protein measurements could be considered for predicting COVID-19 severity, but this requires further study, including more sampling at the onset of symptoms. Moreover, testing data from Mount Sinai did not include sample processing time, which is an important baseline model feature, as well as clinical risk factors. Last, the clinical translation of this study is hindered by the cost involved in measuring 4,701 circulating proteins but could be improved by developing a specific assay to the selected proteins.

4.4 Miscellaneous

4.4.1 Ethics declarations

All contributing cohorts to the present analyses received ethics approval from their respective research ethics review boards. The Biobanque Québécoise de la COVID-19 (BQC19) received ethical approval from the IRB of the JGH and the CHUM. This research was reviewed and approved by the Icahn School of Medicine at Mount Sinai Program for the Protection of Human Subjects (PPHS) under study number 20-00341. This research was considered minimal risk Human Subjects Research.

4.4.2 Data Availability

Code used in this analysis is available at https://github.com/chenyangsu/somalogic with additional information available upon request. The BQC19 is an Open Science Biobank. Instructions on how to access data for individuals from the BQC19 at the Jewish General Hospital site is available at https://www.mcgill.ca/genepi/mcg-covid-19-

biobank. Instructions on how to access data from other sites of the BQC19 is available at https://www.bqc19.ca/en/access-data-samples.

4.4.3 Author Contribution

The author of this thesis was involved in all parts of this work including the conception and design, data analyses, interpretation of data, validation, visualization, and writing, reviewing, revision and editing of this project.

4.4.4 Acknowledgements

The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825; 409511, 100558), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, McGill University, Cancer Research UK [grant number C18281/A29019], and the Fonds de Recherche Québec Santé (FRQS). These funding agencies had no role in the design, implementation, or interpretation of this study. The measurement of proteomics using the SomaLogic panel was supported by the McGill Interdisciplinary Initiative in Infection and Immunity (MI4) and by the Bill and Melinda Gates Foundation for the Mount Sinai cohort.

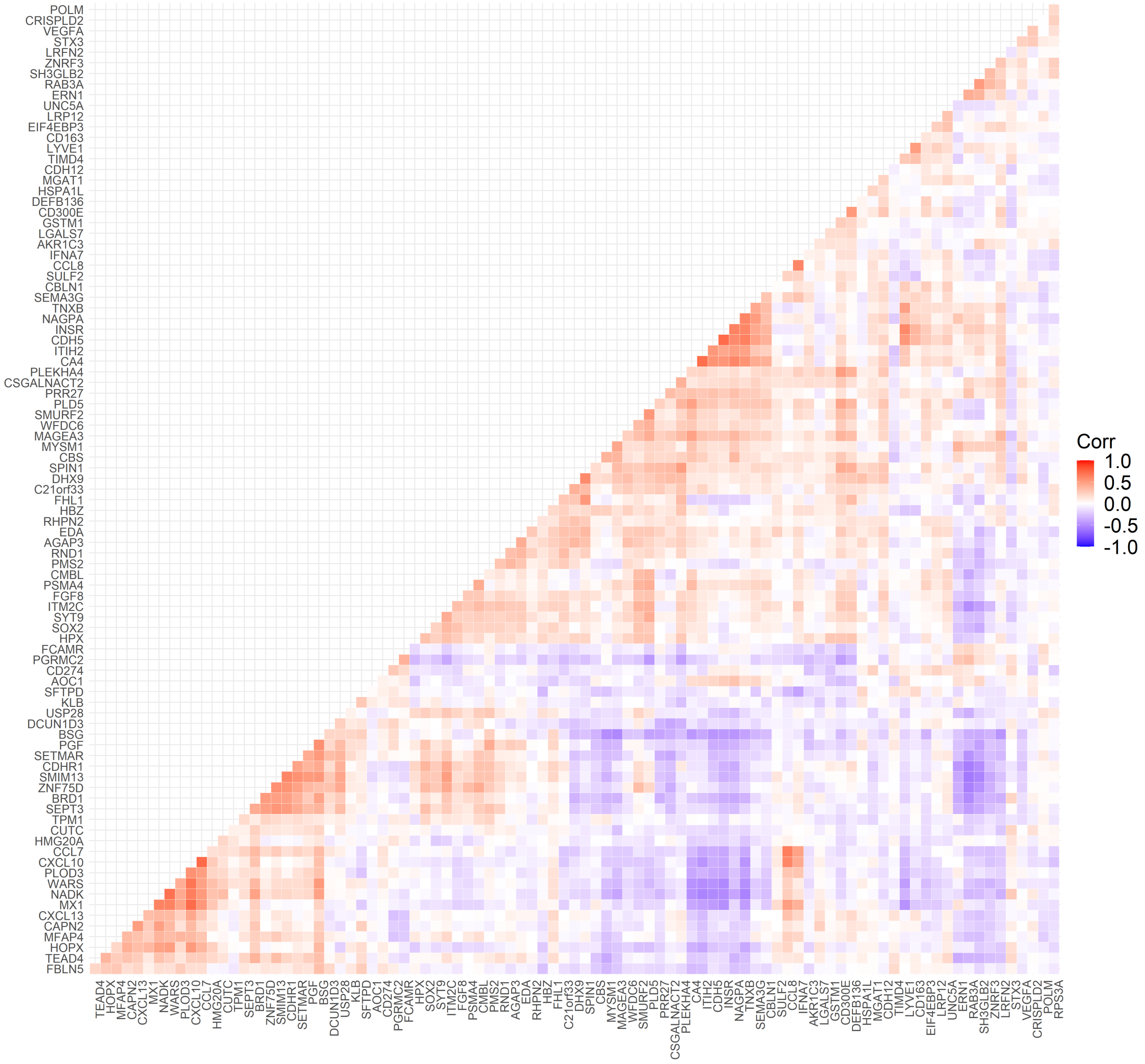


Figure 4.8: Correlations of SOMAmer reagents selected by the protein model when predicting COVID-19 outcomes. Spearman's rank correlations between the 92 proteins within the protein model trained to predict the severe COVID-19 outcome. Plot shows a 92 × 92 heatmap containing 8,464 total correlations with only the bottom portion of the heatmap being shown. Proteins were hierarchically clustered with 8,372 correlations (98.9%) having Spearman's absolute $\rho < 0.8$.

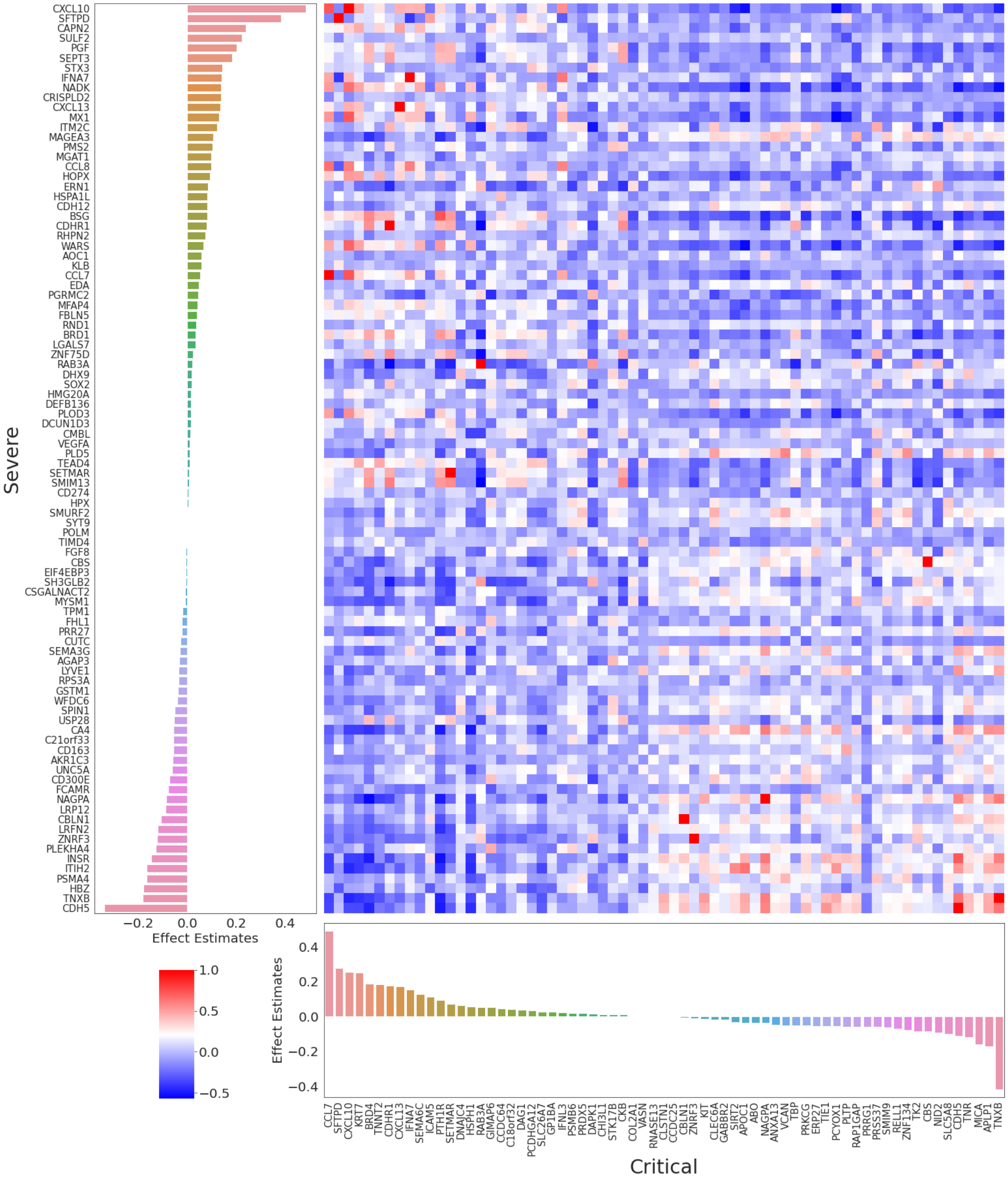


Figure 4.9: Feature importance and correlation of SOMAmers selected in the protein model to predict severe and critical COVID-19. (Rest of caption on next page)

Figure 4.9: (Previous page.) (Left) Coefficient values of the 92 nonzero SOMAmer reagents in the final trained L_1 regularized logistic regression protein model fitted on the severe COVID-19 outcome. The original data contained 4,984 SOMAmer reagents and 4 other variables: age, sex, sample processing time, and hospital site. 92 SOMAmer reagents remained within the model along with age and sample processing time which are not shown. The model was trained on the entire BQC19 cohort using λ = 10.0 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

(Bottom) Coefficient values of the 67 nonzero SOMAmer reagents of the final trained L_1 regularized logistic regression protein model fitted on the critical COVID-19 outcome. The original data contained 4,984 SOMAmer reagents and 4 variables age, sex, sample processing time, and hospital site. 67 SOMAmer reagents remained within the model along with age and sample processing time which are not shown. The model was trained on the entire BQC19 cohort using λ = 10.0 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

(Right) Spearman's rank correlations between the 92 proteins associated with the severe COVID-19 outcome and the 67 proteins associated with the critical COVID-19 outcome. These results suggest that while there were 14 overlapping proteins (SFTPD, CXCL10, RAB3A, NAGPA, CDH5, IFNA7, ZNRF3, CBS, CCL7, SETMAR, TNXB, CDHR1, CXCL13, and CBLN1), in general, the protein levels were uncorrelated with one another. Out of 6,164 total correlations (92 × 67), 6,150 correlations (99.8%) had a Spearman's absolute $\rho < 0.8$.

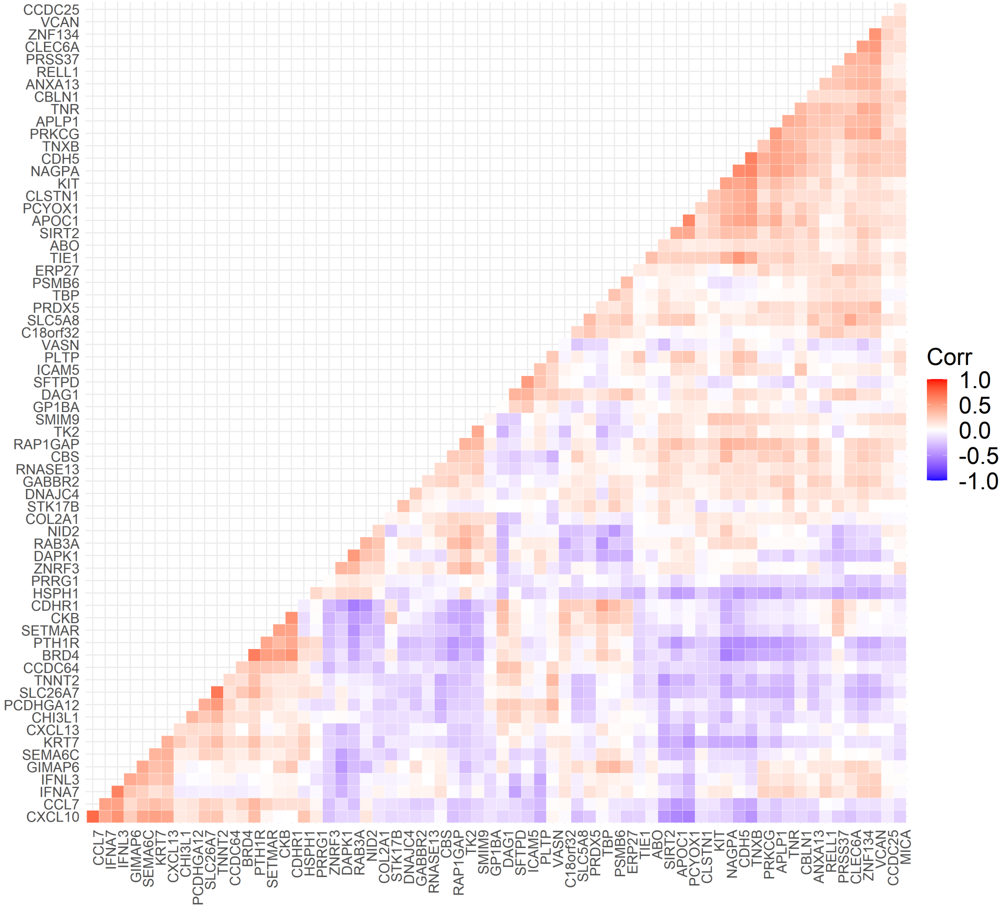


Figure 4.10: Correlations of SOMAmer reagents selected by the protein model when predicting COVID-19 outcomes. Spearman's rank correlations between the 67 proteins within the protein model trained to predict the critical COVID-19 outcome. Plot shows a 67 × 67 heatmap containing 4,489 total correlations with only the bottom portion of the heatmap being shown. Proteins were hierarchically clustered with 4,422 correlations (98.5%) having Spearman's absolute $\rho < 0.8$.

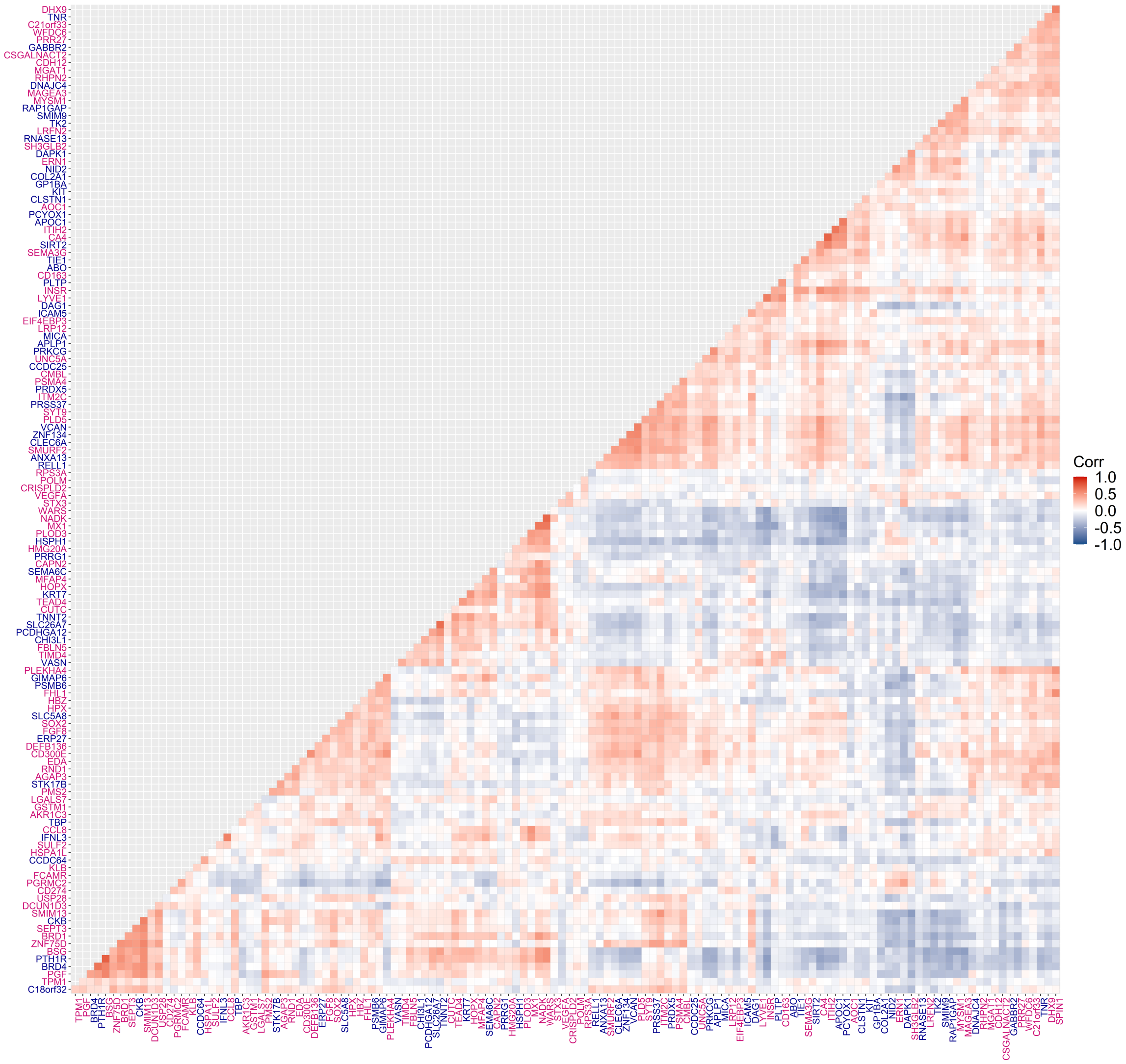


Figure 4.11: Correlations of SOMAmer reagents selected by the protein model when predicting COVID-19 outcomes. Spearman's rank correlations between the nonoverlapping proteins predicting severe (dark blue) and critical (pink) COVID-19 outcomes. Plot shows a 131 x 131 heatmap containing 17,161 total correlations with only the bottom portion of the heatmap being shown. Proteins were hierarchically clustered with 17,030 (99.2%) having Spearman's absolute $\rho < 0.8$.

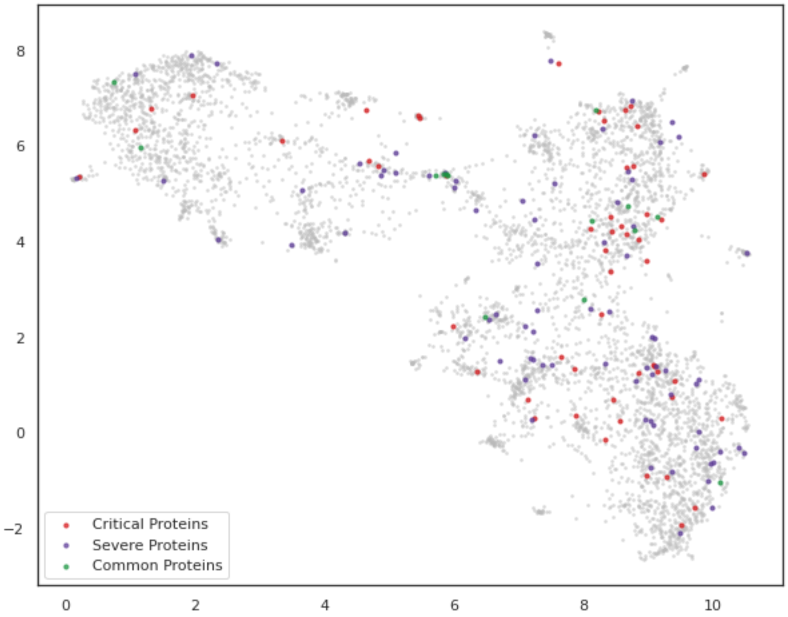


Figure 4.12: Unsupervised clustering by uniform manifold approximation and projection (UMAP) on 4,984 SOMAmer reagents in the BQC19 cohort. Nonlinear dimensionality reduction of measured levels of 4,984 SOMAmer reagents from the BQC19 cohort projected into a 2-dimensional space. Proteins selected by L_1 regularized logistic regression for predicting critical COVID-19 using the protein model are shown in red and proteins selected for predicting severe COVID-19 are shown in purple. The 14 common proteins that were selected in protein models trained on both severe and critical COVID-19 are shown in green.

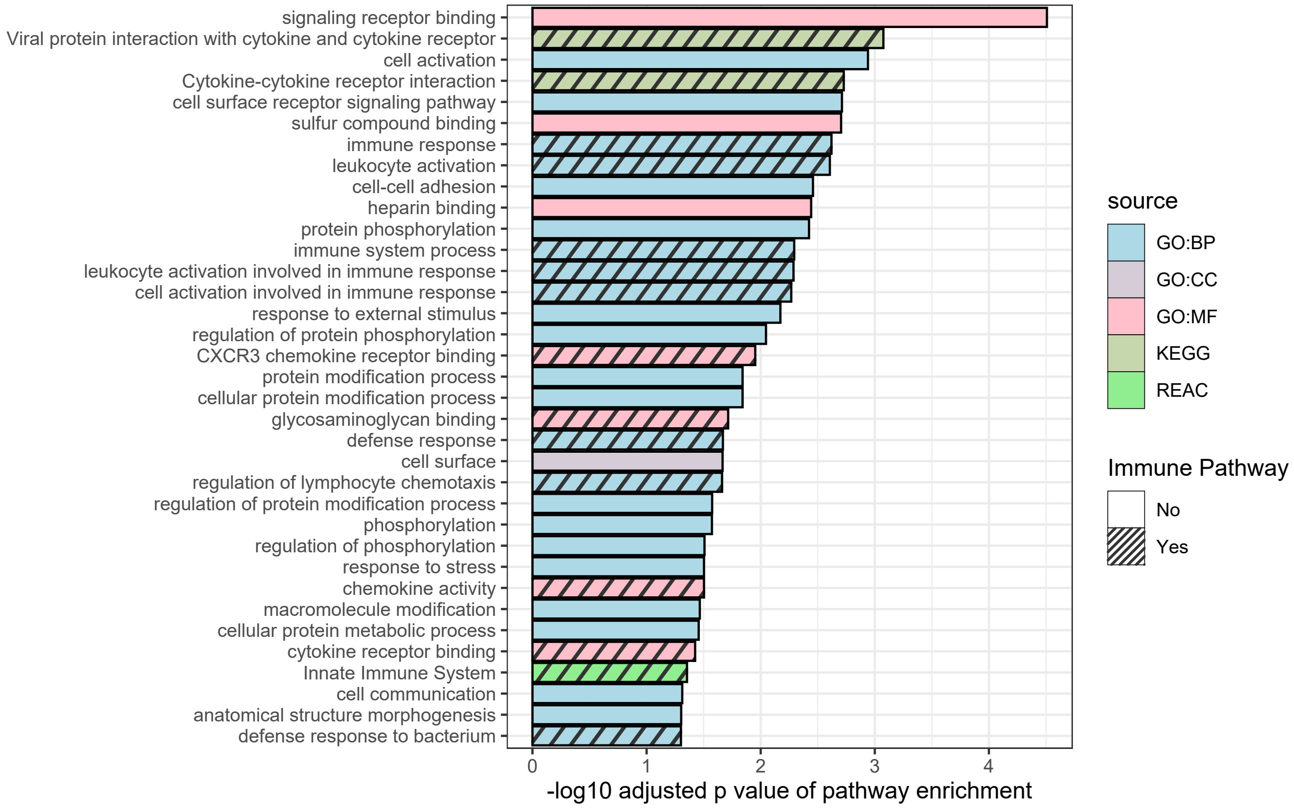
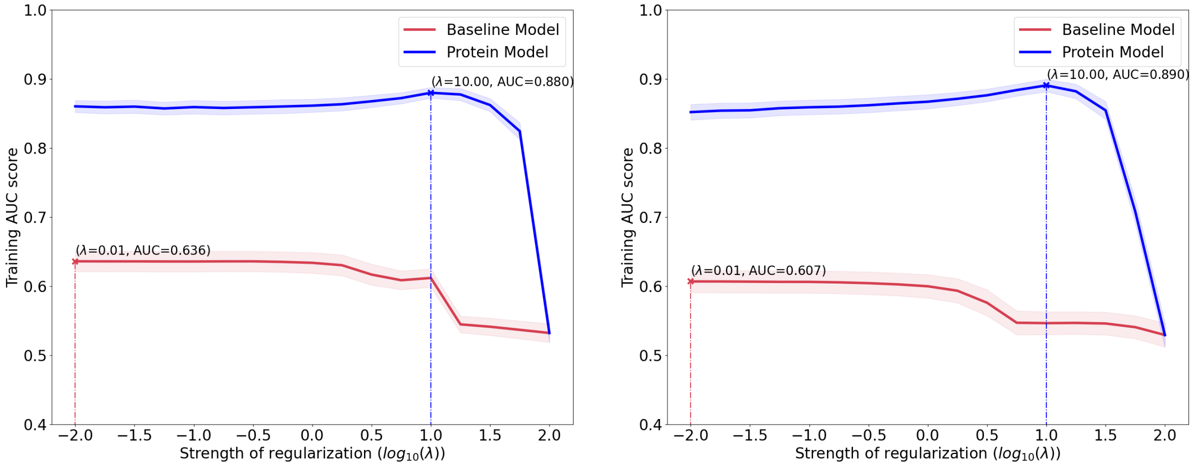


Figure 4.13: Pathway analysis on common proteins best predicting severe and critical COVID-19 outcomes. 32 proteins were common between proteins selected by LASSO that are predictive of critical COVID-19 plus their highly correlated proteins (N=96, Spearman's absolute $\rho > 0.75$) and proteins selected by LASSO that are predictive of severe COVID-19 plus their highly correlated proteins (N=171, Spearman's absolute $\rho > 0.75$). Pathway enrichment was predicted by g:Profiler. Pathways with g:SCS adjusted P < 0.05 are listed. GO_BP: Gene Ontology biological process; GO_CC: Gene Ontology cellular component; GO_MF: Gene Ontology molecular function; KEGG: Kyoto Encyclopedia of Genes and Genomes; REAC: Reactome Pathway Database. Diagonally patterned bars represent immune response pathways.

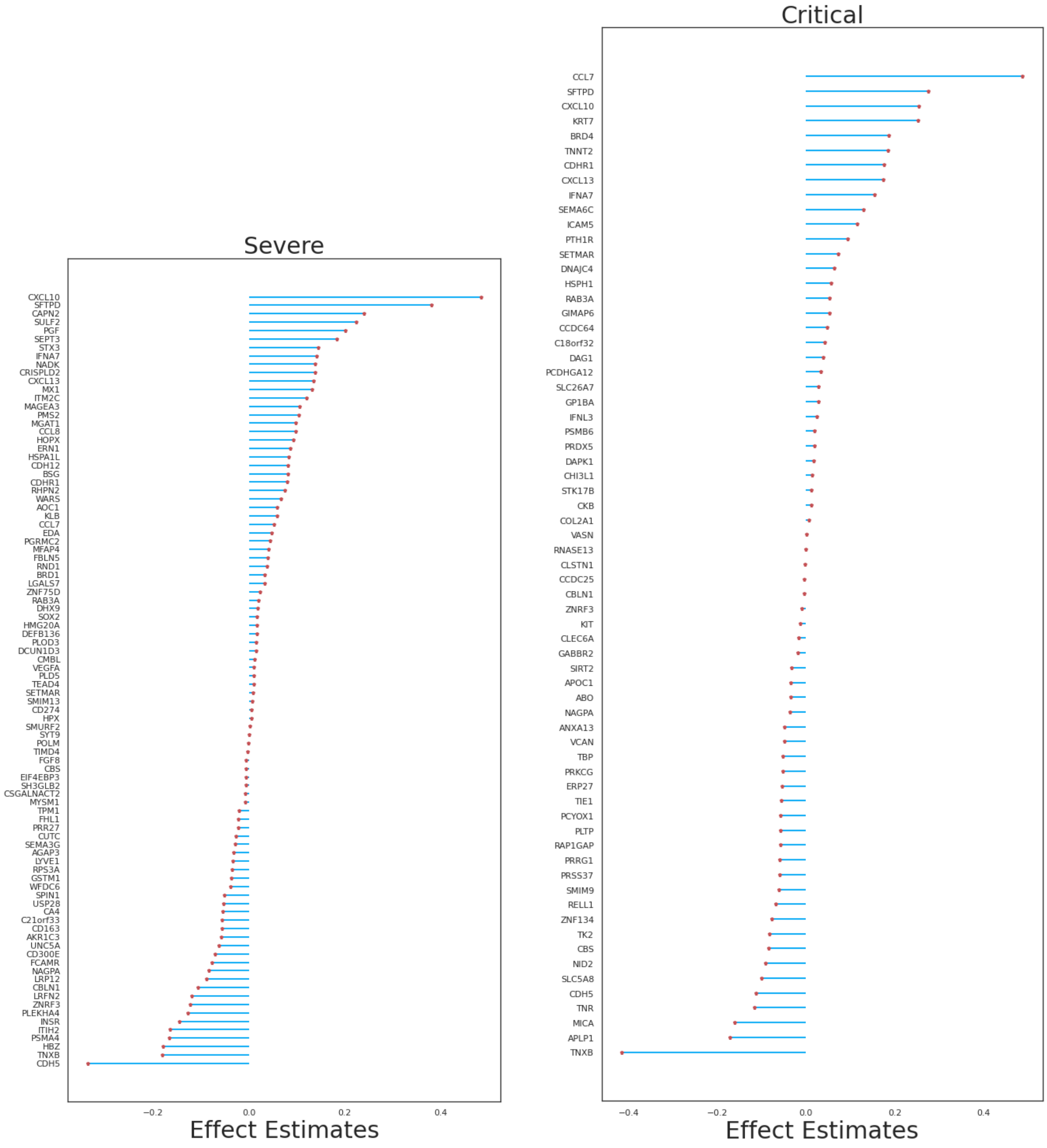


(a) Severe COVID-19

(b) Critical COVID-19

Figure 4.14: Main Analysis plus 6 clinical risk factors. We trained a baseline model (red) with covariates age, sex, sample processing time, hospital site, and 6 clinical features as well as a protein model (blue) containing all baseline model covariates along with an additional 4,984 SOMAmer reagents. The sample size used in this analysis was the entire BQC19 cohort comprising 417 patients.

Train AUC score as a function of regularization strength during L_1 regularized logistic regression when training to predict **a**, severe COVID-19 and **b**, critical COVID-19. The baseline model was fitted with covariates age, sex, sample processing time, hospital site, and 6 clinical features (diabetes, chronic obstructive pulmonary disease, chronic kidney disease, congestive heart failure, hypertension, liver disease). The protein model was fitted with all variables in the baseline model along with 4,984 SOMAmer reagents. Each data point represents an AUC score averaged over 50 validation folds. Shaded areas represent 95% confidence intervals above and below the mean AUC.



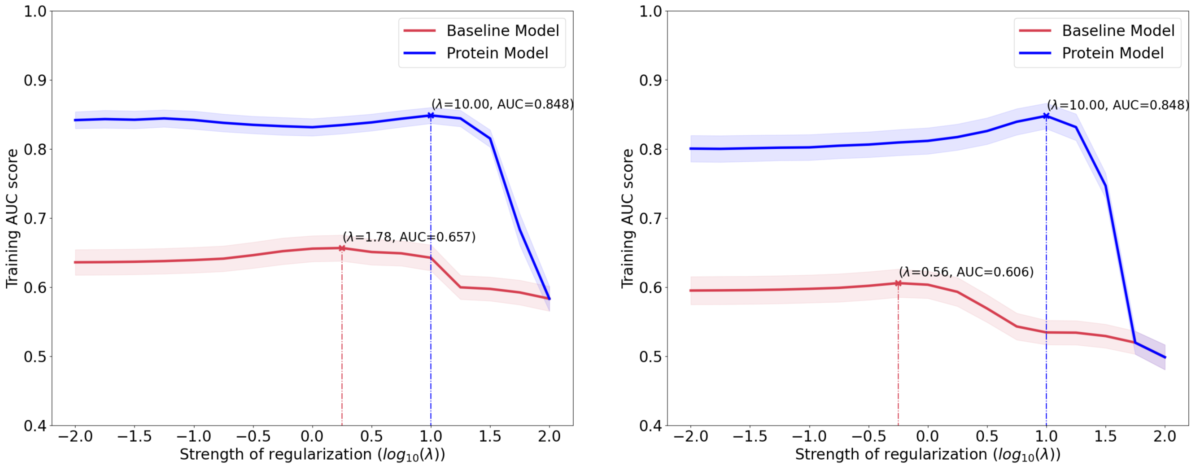
(a) Severe COVID-19

(b) Critical COVID-19

Figure 4.15: Main Analysis plus 6 clinical risk factors: model coefficients. (Rest of caption on next page.)

Figure 4.15: (Previous page.) **a**, Absolute values of the 95 nonzero coefficients of the final trained L_1 regularized logistic regression protein model for predicting severe COVID-19. The original data contained 4,984 proteins and 4 variables age, sex, sample processing time, hospital site, smoking status along with 6 clinical features. A total of 93 proteins remained within the model along with age and sample processing time. The model was trained on the entire training set using λ = 10.0 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

b, Absolute values of the 69 nonzero coefficients of the final trained L_1 regularized logistic regression protein model for predicting critical COVID-19. The original data contained 4,984 proteins and 4 variables age, sex, sample processing time, hospital site, along with 6 clinical features. A total of 67 proteins remained within the model along with age and sample processing time. The model was trained on the entire training set using λ = 10.00 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

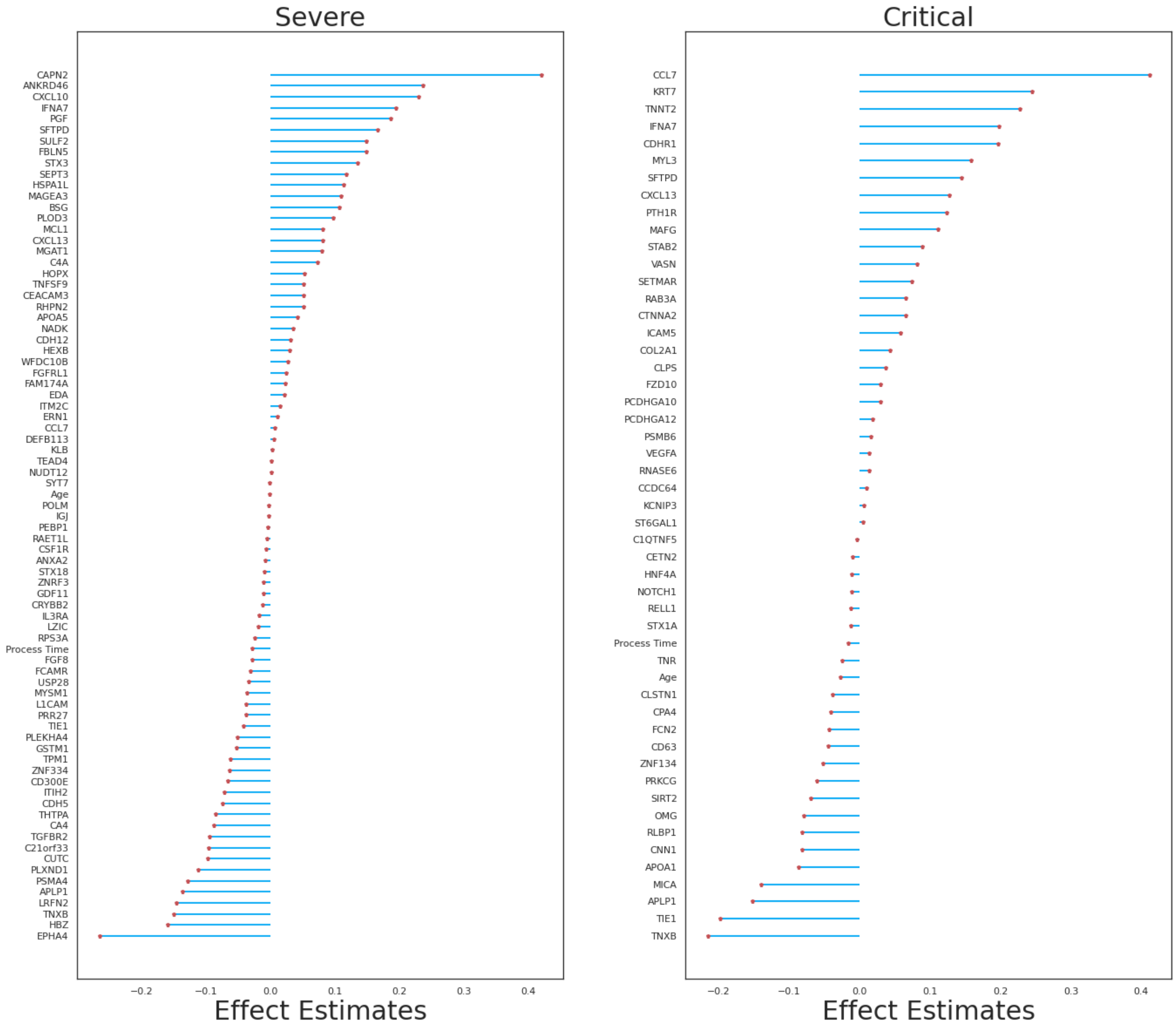


(a) Severe COVID-19

(b) Critical COVID-19

Figure 4.16: Main Analysis plus 7 clinical risk factors. We trained a baseline model (red) with covariates age, sex, sample processing time, hospital site (JGH only), 6 clinical features, and smoking status as well as a protein model (blue) containing all baseline model covariates along with an additional 4,984 SOMAmer reagents. Due to the smoking status variable missing from the CHUM data, the sample size was reduced from 417 samples down to 312 samples for the analysis.

Training AUC score as a function of regularization strength during L_1 regularized logistic regression when trained to predict **a**, severe COVID-19. **b**, critical COVID-19. The baseline model was fitted with covariates age, sex, sample processing time, hospital site, smoking, and 6 clinical features (diabetes, chronic obstructive pulmonary disease, chronic kidney disease, congestive heart failure, hypertension, liver disease). The protein model was fitted with all variables in the baseline model along with 4,984 SOMAmer reagents. Each data point represents an AUC score averaged over 50 validation folds. Shaded areas represent 95% confidence intervals above and below the mean AUC.



(a) Severe COVID-19

(b) Critical COVID-19

Figure 4.17: Main Analysis plus 7 clinical risk factors: model coefficients. (Rest of caption on next page.)

Figure 4.17: (Previous page.) **a**, Absolute values of the 79 nonzero coefficients of the final trained L_1 regularized logistic regression protein model for predicting severe COVID-19. The original data contained 4,984 proteins and 4 variables age, sex, sample processing time, hospital site, smoking status along with 6 clinical features. A total of 77 proteins remained within the model along with age and sample processing time. The model was trained on the entire training set using λ = 10.00 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

b, Absolute values of the 51 nonzero coefficients of the trained L_1 regularized logistic regression protein model for predicting critical COVID-19. The original data contained 4,984 proteins and 4 variables age, sex, sample processing time, hospital site, smoking status along with 6 clinical features. A total of 49 proteins remained within the model along with age and sample processing time. The model was trained on the entire training set using λ = 10.00 (log_{10} λ = 1.0) which was the best λ value found from the hyperparameter search.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main goal of this work was to build predictive models that can improve the outcomes for COVID-19 patients using proteomic data since doing so would provide insights into the causes of severe disease as well as enable the identification of individuals at high risk of severe COVID-19. In this work, we measured 4,701 circulating human protein abundances in two geographically separated cohorts across two countries, totaling 986 individuals. We then trained prediction models including protein abundances and clinical risk factors to predict adverse COVID-19 outcomes in 417 subjects and tested these models in a second, independent cohort of 569 individuals from Mount Sinai. In the most comprehensive large-scale assessment of the role of the proteome in severe COVID-19 outcomes to date, we found that circulating proteins were strongly associated with COVID-19 outcomes and were able to discriminate these outcomes with reasonable accuracy. We also found that protein levels were better able to predict COVID-19 outcomes than nearly all clinical risk factors tested.

In summary, circulating protein levels are strongly associated with COVID-19 outcomes and able to predict the need for oxygen supplementation or death with reasonable accuracy. Measured protein levels were superior to predicting COVID-19 severity outcomes when compared to nearly all clinical risk factors tested. Further research is needed to assess whether this proteomic approach can be applied in a clinical setting to assist in triaging patients for admission to hospital.

5.2 Future Work

The COVID-19 pandemic has presented an unprecedented challenge to economies and public health worldwide. The cooperativeness and teamwork between researchers in interdisciplinary fields including infectious diseases, population genetics, and statistics is crucial to obtain greater insights and treatments for disease. Due to the large collaborative nature of COVID-19 research as well as the impact of the translation of findings that could have on the scientific community and general audiences, quicker and simpler methods for bypassing data sharing agreements may be needed and could be further explored.

The use of federated learning, a machine learning technique which allows training of a model on decentralized data sources may be valuable in improving cooperation and communication between both Canadian and international researchers and improve the speed at which results and findings can be transmitted to benefit not only the scientific community but also the public.

While the goal was to build an efficient and interpretable predictive model to screen for individuals who could potentially develop severe COVID-19, more complex models may be constructed with better predictive performance but at the cost of interpretability. Doing so will improve triage of individuals in the emergency room and highlight individuals who should be monitored more closely for COVID-19 progression to severe outcomes. Further, an understanding of the association patterns and etiological role of these circulating proteins will also ultimately allow for potential protective measures in the form of developmental drug targets regulating key protein expression levels to be discovered, designed, and distributed for use.

More research is needed to determine how variation in the human genome and underlying proteins affect SARS-CoV-2 infection. With the evidence provided by this research work, we will enable evidence-based decisions to be made in selecting molecular targets which can drastically improve the probability of success rates during therapeutic development and lead to better research and development of vaccines and drugs to combat communicable diseases in the future.

Other future work can involve the assessment of how COVID-19 infection affects the genetic regulation of circulating proteins. By diving into genes regulating the expression levels of proteins (pQTLs), this may enable the discovery of underlying biomarkers that can serve as new therapeutic targets to treat COVID-19 infection [110–112]

Bibliography

[1] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, may 2020.

[2] D M Weinreich et al. REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19. 2019.

[3] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. 23(22):5866–5878, 2014.

[4] Sirui Zhou, Guillaume Butler-Laporte, Tomoko Nakanishi, et al. A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nature Medicine*, 27(4):659–667, 2021.

[5] The COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv*, page 2021.03.10.21252820, 2021.

[6] Fei Zhou et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–1062, 2020.

[7] Tomoko Nakanishi et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *medRxiv*, page 2021.03.07.21252875, jan 2021.

[8] Elizabeth J Williamson et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436, 2020.

[9] Yvan Jamilloux, Thomas Henry, Alexandre Belot, Sébastien Viel, Maxime Fauter, Thomas El Jammal, Thierry Walzer, Bruno François, and Pascal Sève. Should we stimulate or suppress immune responses in COVID-19? Cytokine and anti-cytokine interventions. *Autoimmunity Reviews*, 19(7):102567, 2020.

[10] Yufang Shi, Ying Wang, Changshun Shao, Jianan Huang, Jianhe Gan, Xiaoping Huang, Enrico Bucci, Mauro Piacentini, Giuseppe Ippolito, and Gerry Melino. COVID-19 infection: the perspectives on immune responses. *Cell Death & Differentiation*, 27(5):1451–1454, 2020.

[11] Eileen P Scully, Jenna Haverfield, Rebecca L Ursin, Cara Tannenbaum, and Sabra L Klein. Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nature Reviews Immunology*, 20(7):442–447, 2020.

[12] Zhuo Zhou et al. Heightened innate immune responses in the respiratory tract of COVID-19 patients. *Cell Host & Microbe*, 27(6):883–890.e2, 2020.

[13] Jérôme Hadjadj et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science*, 369(6504):718 LP – 724, aug 2020.

[14] Dhiraj Acharya, GuanQun Liu, and Michaela U Gack. Dysregulation of type I interferon responses in COVID-19. *Nature Reviews Immunology*, 20(7):397–398, 2020.

[15] Jeong Seok Lee and Eui-Cheol Shin. The type I interferon response in COVID-19: implications for treatment. *Nature Reviews Immunology*, 20(10):585–586, 2020.

[16] Diane Marie Del Valle et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nature Medicine*, 26(10):1636–1643, 2020.

[17] Zeyu Chen and E John Wherry. T cell responses in patients with COVID-19. *Nature Reviews Immunology*, 20(9):529–536, 2020.

[18] Stephen A. Williams et al. Plasma protein patterns as comprehensive indicators of health. *Nature Medicine*, 25(12):1851–1857, 2019.

[19] Clare Paterson et al. Application of a 27-protein candidate cardiovascular surrogate endpoint to track risk ascendancy and resolution in COVID-19 . 2020.

[20] Peter Ganz et al. Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA - Journal of the American Medical Association*, 315(23):2532–2541, 2016.

[21] Ashok Narasimhan, Safi Shahda, Joshua K Kays, Susan M Perkins, Lijun Cheng, Katheryn N H Schloss, Daniel E I Schloss, Leonidas G Koniaris, and Teresa A Zimmers. Identification of Potential Serum Protein Biomarkers and Pathways for Pancreatic Cancer Cachexia Using an Aptamer-Based Discovery Platform. *Cancers*, 12(12), dec 2020.

[22] Mark Y Chan et al. Prioritizing Candidates of Post-Myocardial Infarction Heart Failure Using Plasma Proteomics and Single-Cell Transcriptomics. *Circulation*, 142(15):1408–1421, oct 2020.

[23] Anne M Lynch, Brandie D Wagner, Alan G Palestine, Nebojsa Janjic, Jennifer L Patnaik, Marc T Mathias, Frank S Siringo, and Naresh Mandava. Plasma Biomarkers of Reticular Pseudodrusen and the Risk of Progression to Advanced Age-Related Macular Degeneration. *Translational vision science & technology*, 9(10):12, sep 2020.

[24] Joseph Yang, Edward N Brody, Ashwin C Murthy, Robert E Mehler, Sophie J Weiss, Robert K DeLisle, Rachel Ostroff, Stephen A Williams, and Peter Ganz. Impact

of Kidney Function on the Blood Proteome and on Protein Cardiovascular Risk Biomarkers in Patients With Stable Coronary Heart Disease. *Journal of the American Heart Association*, 9(15):e016463, aug 2020.

[25] Lasse Folkersen et al. Genomic evaluation of circulating proteins for drug target characterisation and precision medicine. *bioRxiv*, page 2020.04.03.023804, jan 2020.

[26] Andrew L Hopkins and Colin R Groom. The druggable genome., sep 2002.

[27] Tala M Bakheet and Andrew J Doig. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4):451–457, feb 2009.

[28] Martin Lauss, Albert Kriegner, Klemens Vierlinger, and Christa Noehammer. Characterization of the drugged human genome. *Pharmacogenomics*, 8(8):1063–1073, aug 2007.

[29] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12):993–996, 2006.

[30] Lars Wallentin et al. Angiotensin-converting enzyme 2 (ACE2) levels in relation to risk factors for COVID-19 in two large cohorts of patients with atrial fibrillation. *European Heart Journal*, 41(41):4037–4046, nov 2020.

[31] Tom G Richardson, Si Fang, Ruth E Mitchell, Michael V Holmes, and George Davey Smith. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine*, 64:103228, feb 2021.

[32] Maik Pietzner et al. Genetic architecture of host proteins interacting with SARS-CoV-2. *bioRxiv : the preprint server for biology*, jul 2020.

[33] Michael R Filbin et al. Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell reports. Medicine*, 2(5):100287, may 2021.

[34] Prabhu S Arunachalam et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science*, 369(6508):1210 LP – 1220, sep 2020.

[35] Seyma Yasar, Cemil Colak, and Saim Yologlu. Artificial Intelligence-Based Prediction of Covid-19 Severity on the Results of Protein Profiling. *Computer Methods and Programs in Biomedicine*, 202:105996, 2021.

[36] Yapeng Su et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. *Cell*, 183(6):1479–1495.e20, 2020.

[37] Jack Gisby et al. Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. *eLife*, 10:e64827, 2021.

[38] Liis Haljasmägi et al. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. *Scientific Reports*, 10(1):20533, 2020.

[39] Yang Yang et al. Exuberant elevation of IP-10, MCP-3 and IL-1ra during SARS-CoV-2 infection is associated with disease severity and fatal outcome. *medRxiv*, page 2020.03.02.20029975, jan 2020.

[40] Camila Rosat Consiglio et al. The Immunology of Multisystem Inflammatory Syndrome in Children with COVID-19. *Cell*, 183(4):968–981.e7, 2020.

[41] Hamel Patel, Nicholas J Ashton, Richard J B Dobson, Lars-Magnus Andersson, Aylin Yilmaz, Kaj Blennow, Magnus Gisslen, and Henrik Zetterberg. Proteomic blood profiling in mild, severe and critical COVID-19 patients. *Scientific Reports*, 11(1):6357, 2021.

[42] Karine Tremblay 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):e0245031, may 2021.

[43] Technical Note. SomaScan ® Assay v4.0 Technical Note. 2021.

[44] Haitao Ma, Jinping Liu, M Monsur Ali, M Arif Iftakher Mahmood, Louai Labanieh, Mengrou Lu, Samir M Iqbal, Qun Zhang, Weian Zhao, and Yuan Wan. Nucleic acid aptamers in cancer research, diagnosis and therapy. *Chemical Society reviews*, 44(5):1240–1256, mar 2015.

[45] Alexander W. Charney et al. Sampling the host response to SARS-CoV-2 in hospitals under siege. *Nature Medicine*, 26(8):1157–1158, 2020.

[46] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, jan 2010.

[47] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association : JAMIA*, 21(4):699–706, 2014.

[48] Maja Buszko, Aleksandra Nita-Lazar, Jung-Hyun Park, Pamela L Schwartzberg, Daniela Verthelyi, Howard A Young, and Amy S Rosenberg. Lessons learned: new insights on the role of cytokines in COVID-19. *Nature Immunology*, 22(4):404–411, 2021.

[49] Víctor J Costela-Ruiz, Rebeca Illescas-Montes, Jose M Puerta-Puerta, Concepción Ruiz, and Lucia Melguizo-Rodríguez. SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine & growth factor reviews*, 54:62–75, aug 2020.

[50] Daniel E Leisman et al. Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. *The Lancet Respiratory Medicine*, 8(12):1233–1244, dec 2020.

[51] David C Fajgenbaum and Carl H June. Cytokine Storm. *New England Journal of Medicine*, 383(23):2255–2273, dec 2020.

[52] Yingxia Liu et al. Elevated plasma levels of selective cytokines in COVID-19 patients reflect viral load and lung injury. *National Science Review*, 7(6):1003–1011, jun 2020.

[53] Sara De Biasi et al. Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nature Communications*, 11(1):3434, 2020.

[54] Adam G Laing et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nature Medicine*, 26(10):1623–1635, 2020.

[55] Amber L Mueller, Maeve S McNamara, and David A Sinclair. Why does COVID-19 disproportionately affect older people? *Aging*, 12(10):9959–9981, may 2020.

[56] Monica Fung and Jennifer M Babik. COVID-19 in Immunocompromised Hosts: What We Know So Far. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 72(2):340–350, jan 2021.

[57] Rachel M Ostroff et al. Unlocking Biomarker Discovery: Large Scale Application of Aptamer Proteomic Technology for Early Detection of Lung Cancer. *PLOS ONE*, 5(12):e15003, dec 2010.

[58] Rachel M Ostroff et al. Early Detection of Malignant Pleural Mesothelioma in Asbestos-Exposed Individuals with a Noninvasive Proteomics-Based Surveillance Tool. *PLOS ONE*, 7(10):e46091, oct 2012.

[59] Heli Julkunen et al. Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *eLife*, 10:e63033, 2021.

[60] Wentao Ni et al. Role of angiotensin-converting enzyme 2 (ACE2) in COVID-19. *Critical Care*, 24(1):422, 2020.

[61] Felix W Frueh and Michael E Burczynski. Chapter 37 - Large-scale molecular profiling approaches facilitating translational medicine: genomics, transcriptomics, proteomics, and metabolomics. pages 699–718. Academic Press, 2021.

[62] William S Bush and Jason H Moore. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, 8(12):e1002822, dec 2012.

[63] Yukihide Momozawa and Keijiro Mizukami. Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, 66(1):11–23, Jan 2021.

[64] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

[65] Yosef Hochberg and Ajit C. Tamhane, editors. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., September 1987.

[66] Hervé Abdi et al. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107, 2007.

[67] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[68] Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[69] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[70] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, nov 2018.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[72] Sorin Mihai Grigorescu, Bogdan Trasnea, Tiberiu T. Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *CoRR*, abs/1910.07738, 2019.

[73] Vaishnavi Nath Dornadula and S Geetha. Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, 165:631–641, 2019. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[74] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, feb 2019.

[75] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey, 2019.

[76] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[77] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[78] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[79] Léon Bottou. On-line learning and stochastic approximations. In *In On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press, 1998.

[80] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[81] Robert A. McLean, William L. Sanders, and Walter W. Stroup. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64, 1991.

[82] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[83] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 5. pearson Boston, MA, 2007.

[84] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[85] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, may 1996.

[86] A.Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[87] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[88] Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 05 2005.

[89] Panagiotis (Panos) Adamopoulos. https://netman.aiops.org/∼peidan/anm2018/3.machinelearn

[90] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89, 1997.

[91] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[92] David W. Hosmer and Stanley Lemeshow. *Applied logistic regression*. John Wiley and Sons, 2000.

[93] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[94] W J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, jan 1950.

[95] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[96] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[97] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[98] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[99] M Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, may 2000.

[100] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000.

[101] Bijay Jassal et al. The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503, jan 2020.

[102] E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. *Nucleic Acids Research*, 24(1):238–241, 01 1996.

[103] Sheng-Da Hsu et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic acids research*, 39(Database issue):D163–D169, jan 2011.

[104] Mathias Uhlén et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[105] Madalina Giurgiu et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*, 47(D1):D559–D563, 10 2018.

[106] Certain Medical Conditions and Risk for Severe COVID-19 Illness — CDC.

[107] Wenhua Liang et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA internal medicine*, 180(8):1081–1089, aug 2020.

[108] Stephen R. Knight et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *The BMJ*, 370(September):1–13, 2020.

[109] Gareth J. Griffith et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications*, 11(1):1–12, 2020.

[110] Benjamin B. Sun et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[111] Liam Gaziano et al. randomization identifies repurposing opportunities for COVID-19. 1.

[112] Lucija Klaric et al. Mendelian randomisation identifies alternative splicing of the FAS death receptor as a mediator of severe COVID-19. *medRxiv : the preprint server for health sciences*, page 2021.04.01.21254789, apr 2021.

Appendix A

Supplementary Tables

Supplementary tables are attached as a separate file called **SupplementaryTables.xlsx**. Below are the descriptions of each supplementary table.

- Supplementary Table 1: Logistic regression of a single SOMAmer reagent on severe COVID-19, adjusting for sex, age, sample processing time, and hospital site.
- Supplementary Table 2: Logistic regression of a single SOMAmer reagent on critical COVID-19, adjusting for sex, age, sample processing time, and hospital site.
- Supplementary Table 3. Betas of 92 non-zero proteins identified by lasso that were associated with severe COVID-19.
- Supplementary Table 4. Beta of 67 non-zero proteins identified by lasso that were associated with critical COVID-19.
- Supplementary Table 5. Proteins with Spearman's absolute $\rho > 0.75$ when correlated with one of the proteins in ST3
- Supplementary Table 6. Proteins with Spearman's absolute $\rho > 0.75$ when correlated with one of the proteins in ST4
- Supplementary Table 7. Pathway enrichment of 32 common proteins

- Supplementary Table 8. Mount Sinai COVID-19 Biobank Team