Instituto Politécnico de Lisboa

Instituto Superior de Engenharia De Lisboa

Escola Superior de Tecnologia Da Saúde De Lisboa

Biomarkers Discovery For Prognosis of COVID-19 Based on Metabolomics

Tiago Alexandre Henrique Fonseca

Thesis to obtain the Master degree in **Biomedical Engineering**

Supervisors:

Cecília R.C. Calado (ISEL – Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa)

Conceição Oliveira (Instituto Superior Técnico, Universidade de Lisboa)

Luís Bento (Centro Hospitalar Universitário Lisboa Central)

Examination Committee:

Chairperson

Miguel Minhalma (ISEL)

Members of the Committee

Ana Fernandes Platzgummer (ISEL)

Cecília R.C. Calado (ISEL)

June 2022

Instituto Politécnico de Lisboa

Instituto Superior de Engenharia de Lisboa

Escola Superior de Tecnologia da Saúde de Lisboa

Biomarkers Discovery For Prognosis of COVID-19 Based on Metabolomics

Tiago Alexandre Henrique Fonseca

Thesis to obtain the Master degree in **Biomedical Engineering**

Supervisors:

Cecília R.C. Calado (ISEL – Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa)

Conceição Oliveira (Instituto Superior Técnico, Universidade de Lisboa)

Luís Bento (Centro Hospitalar Universitário Lisboa Central)

Examination Committee:

Chairperson

Miguel Minhalma (ISEL)

Members of the Committee

Ana Fernandes Platzgummer (ISEL)

Cecília R.C. Calado (ISEL)

June 2022

[Blank Page]

Biomarkers Discovery For Prognosis of COVID-19 Based on Metabolomics

Tiago Alexandre Henrique Fonseca

This work was supported by *Instituto Politécnico de Lisboa* grant IDI&CA/IPL/2020/NephoMD/ISEL, and the FCT grant DSAIPA/DS/0117/2020 - PREMO - Predictive Models of COVID-19 Outcomes for Higher Risk Patients Towards a Precision Medicine.

[Blank Page]

Acknowledgments

Thank you to my supervisor, Professor Dr Cecília Calado, for providing the opportunity to be part of your laboratory team and be under your guidance throughout the thesis and this project, without you I would not be here developing this work, and for this I thank you. Hope this team that we are never separates and continues to produce great works with great results and major impact in our society.

I would like to thank Dr Luís Bento for all the aid and inspiration for this project, without it this thesis would end here, I thank you for the opportunity to be part of this project and hope to work with you in the following projects that are in certain yet to come, and so I thank you.

A major thank to the collaboration with Instituto Superior Técnico, especially for Dr Conceição Oliveira for introducing me to the equipment that delivered me the data I needed for this work. I would like to thank as well Dr Gonçalo Justino for all the help with the metabolomic data interpretation and for giving me such a precious guidance and knowledge throughout the writing process, your patience for all my questions and doubts is a great virtue, thank you for all the help you could give.

I would like to thank Dr António Ricardo for extending his evenings of work with me in the hospital collecting the missing data of patient's database, without you I couldn't have proceeded with this work, thank you.

Engineer Rúben Araújo, where would I be without your guidance and advice, maybe still trying to learn how to do machine learning without Orange. Thank you for the good lab talks that sent me in the right direction and the zoom calls that explained me so much about a major part of this thesis, machine learning. You opened a new path in my academic perspectives, thank you good colleague and hope friend.

Thank you to the future Master, Cristiana Von Rekowski, thank you first for being the love of my life and be there as the best companion as one could ever hope to get and secondly thank you for boarding with me in this journey of complete a thesis. Without you by my side I would be struggling in the statistics of this work and bashing my head against the walls at home. It is a pleasure to share my life and work with you, thank you.

I would like to mention my mother and father all the push for continuing my academic life and giving the advice to not stop with a simple graduation, and to not stop even here. Thank you for always giving me the opportunity to go further in my academic life, without you both this journey would be even harder. As the first member of the family to enter the university and finishing a Master thesis and aiming for a PhD, I watch to you brother, not to follow the same steps but to be better than me, I hope you follow your dreams and be happy with a little cash on your pockets, study hard and live well.

Finally, I would like to thank me, without me I would not be where I am, thank you for not quitting when you wanted to and to sacrifice things in life to get other, thank you for going to work while you studied it only helped you to see the other side of the "real" life. You fought and are still fighting for your life objectives, don't go soft now, keep it up the good work.

Once again,

Thank you all for the opportunities and experience that you gave me.

[Blank Page]

Abstract

Background and Goals: A novel coronavirus strain, SARS-CoV-2, emerged in late 2019, generating a viral epidemic. This new, highly transmissible strain outnumbered both SARS and MERS in terms of affected people. Symptoms of the novel virus included fever, cough, and chest pain, as well as dyspnea and bilateral lung infiltration in severe instances. Due to the relevance of the COVID pandemic, this thesis aims to develop a predictive model for the outcome of critically ill patients based on a metabolomic serum analysis, namely Fourier-transform infrared spectroscopy and mass spectrometry.

Methods: Two assay groups were analysed. The first experiment examined the influence of two distinct extraction techniques for metabolite extraction, as well as a prediction model for the patients' outcome (n=6) using serum data. The second test used a larger sample size (n=24) and examined metabolite expression based on the patient's condition (discharged, deceased, and invasive mechanical ventilation), along with the development of an outcome prediction model, using metabolite analysis. This work was carried out applying FTIR and LC-MS systems, as well as multivariate analytic software.

Results: The findings of the first assay demonstrated that utilising methanol as a solvent for metabolite extraction resulted in a larger lipidic profile in the samples as compared to the methodology employing an acetonitrile:methanol:water solvent mixture, which resulted in a higher peptide output. With an Area Under the Curve of 0.98 and a Classification Accuracy of 0.97, machine learning algorithms predicted the patients' prognosis. In the second experiment, just the acetonitrile:methanol:water procedure was used, and the metabolite and putative metabolism profile of each group was deduced from mass spectrometry data. Classification accuracies of 1.00 were achieved for prediction of patient ventilation and outcome between non-ventilated and deceased ventilated patients, and 0.97 for outcome prediction between ventilated discharged patients and deceased ventilated patients.

Conclusions: Irrespective from the positive results obtained with the algorithms for predicting patient outcomes, it is crucial to note that the study samples were quite small. As a conclusion, further research is necessary to confirm the results of this study.

Keywords: COVID-19; Fourier Transform Infrared Spectroscopy; Mass Spectrometry; Liquid Chromatography; Metabolomics; Principal Component Analysis; Biomarkers; Machine Learning

Resumo

Introdução e Objetivos: Uma nova estirpe de coronavírus, SARS-CoV-2, surgiu no final de 2019, o que gerou um surto pandémico. Esta nova variante altamente transmissível superou tanto a SARS quanto a MERS em termos de pessoas infetadas. Os sintomas deste novo vírus incluem febre, tosse e dor torácica, assim como dispneia e infiltração pulmonar bilateral em casos graves. Devido à relevância da pandemia COVID, esta tese tem como objetivo desenvolver um modelo preditivo para o prognóstico de doentes críticos com base em análise metabolómica de soro nomeadamente espectroscopia de invravermelho por transformada de Fourier e espectrometria de massa.

Métodos: Dois ensaios foram analisados. A primeira análise verificou o impacto de duas técnicas de extração distintas para extração de metabolitos, bem como um modelo de previsão para o desfecho clínico dos doentes (n=6). O segundo teste usou uma amostra maior (n=24) e examinou a expressão de metabolitos com base na condição do paciente (alta, falecimento e ventilação mecânica invasiva), juntamente com o desenvolvimento de um modelo de previsão do desfecho clínico. Este trabalho foi realizado com recurso a sistemas como o FTIR e LC-MS bem como software de análise multivariada.

Resultados: Os resultados do primeiro ensaio demonstraram que a utilização de metanol como solvente para extração de metabolitos conferiu um perfil mais lipídico às amostras em comparação com a extração com acetonitrilo:metanol:água que resultou numa amostra com um perfil mais peptídico. Os modelos de previsão obtiveram resultados de área debaixo da curva de 0.98 e precisão de classificação de 0.98 para o desfecho dos doentes. No segundo ensaio foi utilizado apenas o protocolo acetonitrilo:metanol:água, e os metabolitos e suposto metabolismo de cada grupo foram identificados por espetrometria de massa. Precisões de classificação de 1.00 para o desfecho entre doentes não ventilados e ventilados ambos com alta e entre não ventilados com alta e ventilados falecidos, e de 0.97 entre ventilados com alta e ventilados falecidos para a predição do desfecho clínico foram alcançadas nas comparações entre grupos.

Conclusões: Independentemente dos resultados promissores dos algoritmos de predição, é crucial afirmar que as amostras de estudo são pequenas. Como conclusão, mais pesquisas com tamanhos amostrais superiores são necessárias para corroborar os resultados obtidos.

Palavras-chave: COVID-19; Espectroscopia de Infravermelho com Transformada de Fourier; Espectrometria de massa; Cromatografia Líquida; Metabolómica; Análise de Componente Principal; Biomarcadores; Aprendizagem Automática.

Previous Publications in The Field

Araújo R, Bento LFN, Fonseca TAH, Von Rekowski CP, da Cunha BR, Calado CRC. Infection Biomarkers Based on Metabolomics. Metabolites. 2022 Jan 19;12(2):92. [1]

[Blank Page]

Index

Acknowledgments vi
Abstract viii
Resumo ix
Previous Publications in The Field x
Index xii
List of Figures xiii
List of Tables xix
List of Abbreviations xxi
Chapter 1 1
1.1. Introduction 1
1.2. Objectives 1
Chapter 2 - Literature Review 2
2.1. SARS-CoV-2 Taxonomy, Characterization, and Replication 2
2.2. SARS-CoV-2 Transmission and Pathogenicity 7
2.3. Metabolomics 10
2.4. Non-Conventional Metabolomics 15
Chapter 3 – Materials and Methods 17
3.1. Biological data acquisition 17
3.2. FTIR Spectroscopic analysis 18
3.3. UHPLC-MS Analysis 23
3.4. Patient Groups Assembly 25
Chapter 4 – Results & Discussion 28
4.1. FTIR Analysis Results 28
4.1.1. First Assay Results 28
4.1.2. Second Assay Results 33
4.2. UHPLC-MS Analysis Results 40
4.2.1. First Assay Results – Extraction Procedure Comparison 40
4.2.2. Second Assay Results – Metabolome Analysis for Metabolite Identification 44
Chapter 5 – Conclusion 51
Chapter 6 – Future Work & Perspectives 52
References 53
Appendix 62

List of Figures

Figure 2.1.1 - Evolution model of the *CoV* species. Adapted from Woo *et al*. [13]. 2
Figure 2.1.2 - Taxonomy of coronaviruses compared to the human species. Adapted from Gorbalenya *et al*. [8]. 2
Figure 2.1.3 - Schematic representation of the structure of SARS-CoV-2. Figure 2.1.3 shows an unreported structure, Hemagglutinin-esterase (He) is a glycoprotein existing in some viruses in the envelope zone, which helps in the cell invasion mechanism. However, this molecule is not always identified in the structural composition of the SARS-CoV-2 species. (Image: Orpheus FX/Shutterstock)..... 3
Figure 2.1.4 – Frameshifting process. The three pill-like shaped figure represents the ribosomal complex, while the single stranded RNA template is represented by the line of letters, some of which are red coloured. The tRNA here transports 2 distinct amino acids, asparagine (Asn) and leucine (Leu). 3
Figure 2.1.5 - SARS-CoV-2 genomic structure. composed of ssRNA+ with a cap (grey circle) at the 5' end and a poly-A tail at the 3' end. It encodes 16 nsp's (ORF1a → nsp1-11 and ORF1B → nsp12-16). Structural proteins are identified as S, E, M and N and accessory proteins are numbered and identified by gray rectangles. Adapted from Brant *et al*. [18] 4
Figure 2.1.6 - Spike protein role in host cell invasion (image: AIVD Biotech Inc.). 5
Figure 2.1.7 – SARS-CoV-2 life cycle. Adapted from Burmer *et al*. [43]..... 5
Figure 2.2.1 - COVID-19 pneumonia radiography and CT scans. (A) X-ray image showing a more prominent bilateral pulmonary infiltrate in the left lung. (B) Image obtained by computed tomography showing areas of increased opacity, islets of opacity are bilaterally distributed in the pulmonary zone. Adapted from [22] and [74]. 7
Figure 2.2.2 - Clinical features of COVID-19. COVID-19 clinical characteristics can easily be mistaken for a typical cold or pneumonia when the disease is set on a mild to severe case. Healthy patients tend to be asymptomatic, being the age a heavy risk factor. Age, and age associated co-morbidities are the major factors that lead to higher risk of critical development and in the worse cases, to death. Adapted from Hu *et al* [5]. 8
Figure 2.2.3 – Inflammasome complex formation. It's created during the pathogen-associated molecules (PAMP's) process, such in the case of SARS-CoV-2, and molecules associated with cell damage (DAMPs). PAMPs and DAMPs are recognized at the PRR (pattern recognition receptor) from APCs (Antigen-Presenting Cells). From the coupling between DAMP/PAMP-PRR an inflammatory cascade starts and activates the secretion of pro-inflammatory cytokines (like pro-IL-1β). In the case of the host cell, cytosolic receptors such as Nod-like receptors induce a specialized behavior of programmed cell death, apoptosis, releasing even more antigens capable of triggering a greater immune response. In normal conditions, these processes are regulated by the organism towards beneficial results, however, SARS-CoV-2 manages to dysregulate it. Adapted from Soy *et al* [80]. 8
Figure 2.2.4 - Schematic diagram of the renin-angiotensin system (RAS). In the lung where the role of ACE2 receptor and ACE is visible as regulatory elements of RAS. The RAS pathway begins with the conversion of angiotensinogen to the angiotensin I (AngI) precursor through the action of the enzyme renin, produced in the kidneys. AngI is further recognized by the angiotensin-converting enzyme (ACE) cleaving it to angiotensin II (AngII). AngII is the peptide with the greatest influence on RAS. Its effect are mainly at the level of AT1R receptors. It also couples to the AT2R having the antagonistic effect. On the other hand, ACE2 receptors convert AngI and II into the inactive precursor Angiotensin (1-7) which, when interacting with the MAS receptor, has antagonistic effects to that of AngII [79,81]. 9
Figure 2.2.5 - Pathogenesis caused by SARS-CoV-2 viral infection in the renin-angiotensin system. The virion complexes with the ACE2 membrane receptor. After cell invasion, the natural antiviral pathway is activated, promoting the cellular secretion of interferons that have an apocrine effect with adjacent cells, promoting their expression of ACE2, creating new cellular targets for SARS-CoV-2. However, the levels of available ACE2 go down by sequestration, the levels of angiotensin II are promoted causing an uncontrolled inflammatory reaction with harmful effects at the level of the affected organ. Adapted from Gao *et al*. [81]..... 9
Figure 2.3.1 – Metabolomics work-flow. A typical metabolomic experiment involves several steps: cells, tissues and other biofluids can be isolated and extracted. Samples are then prepared for analysis. Commonly analytical platforms are coupled to other devises such chromatographic platforms, to separate the different components of the sample, reducing the sample complexity and rearrange the major sample molecules, facilitating the MS or NMR job to identify molecular profiles. The data retrieved can be then compared to large databases. After that, statistical and multivariate algorithms are used to verify the significance of those molecules, and then the

molecule(s) that pass through the sorting are tested as a differential metabolite(s), which is, the ability to distinguish normal individuals from patients with the variable of study for example. Adapted from Alseekh *et al* [94] and Li *et al* [87]. 11
Figure 2.4.1 – Every molecule has several modes which they can vibrate at. (A) A nonlinear molecule such as H_2O has a 3N-6 different ways that it can vibrate (being N the number of atom nuclei present on the molecule). (B) linear molecules such as CO_2 has 3N-5 (being N the number of nuclei atoms on the molecule). The vibrational modes pass through stretching, symmetric or antisymmetric, depending on if the O-H bond are in or out of phase, respectively. It can be as well a bending vibration, if the angle between the H-O-H is changing, also called the scissoring motion. In the case of the molecule of CO_2, it has 4 normal modes of vibrations (name given to each way which a molecule can vibrate), symmetrical stretch where the O-C-O bond elongates in opposite directions, maintaining the 180º angle, antisymmetrical stretch where one O-C bond is shorter than the other, still maintain the 180º angle. And finally, the Bend, where the O-C-O 180º angle is changed. Adapted from [106]..... 15
Figure 2.4.2 – Diagram of an inside FTIR spectrometer. The IR light is directed from the incandescent source to a beam splitter, positioned at a 45° angle. A portion of the beam is reflected at a straight angle after passing through the beam splitter. The reflected beam hits a fixed mirror, whereas the transmitted beam hits a moving mirror at a constant speed. The beams recombine at the beam splitter after returning from the mirrors. The moving mirror's beam has travelled a different distance than the fixed mirror's beam, and the two beams combine to form an interference pattern known as an interferogram. This interferogram, which contains all frequencies at the same time, travels through the sample compartment to reach the detector. A Fourier transform, a typical computer procedure, changes the time domain spectrum to the frequency domain spectrum, allowing us to measure the degree of absorption as a function of wavelength (or frequency). Adapt from Wade *et al* [107]. 15
Figure 3.1.1 - Properly identified peripheral blood collection tubes (censored for ethical reasons). 17
Figure 3.1.2 - Stored serum aliquots. 17
Figure 3.1.3 - Schematic view of the procedure executed. The blood from the patients is collected directly from the ICU stored in collection-blood tubes without anticoagulant to facilitate the blood clotting process in which is easier to extract the serum through centrifugation. The serum is aliquoted in at least 2 microtubes with a volume of 2 mL and stored each volume at -20 ºC and -80 ºC refrigerator for further analysis (FTIR and/or LC-MC).. 18
Figure 3.2.1 - ZnSe microplates used for this technique. (A) 96 well ZnSe microplate, with each well holding a volume capacity of 25 µL; (B) 384 well ZnSe microplate with a volume capacity of 5 µL each well. 18
Figure 3.2.2 – Microplate with loaded samples in the wells placed on a rack in a desiccator. In the image is possible to observe two cylinders containers with silica crystals that helps the evaporation process. 19
Figure 3.2.3 - Workflow process done on FTIR analysis. Dried biological samples are read at the FTIR spectrometer. RAW data is loaded into OPUS software for primary pre-processing. Signal normalization and/or derivation is carried out on The Unscrambler® X where the data is smoothed and treated into more amenable signals. 19
Figure 3.2.4 - Spectra obtained from FTIR spectrometer from the solvent extraction assay (explained in chapter 3.4). Spectra in the left present absorption intensity vs. wavenumber, as recorded on the HTS-XT FTIR. (A) Raw spectra with no pre-processing. (C) Spectra submitted to Atmospheric correction executed in OPUS software. (E) Spectra submitted to Atmospheric Correction followed by Baseline, Rubber band method, correction executed on OPUS software. B, D and F are the PCA's of spectra A, C and D respectively, and presented to facilitate a visual insight of the spectra shown. Each dot in the PCA represents a sample, extracted using methanol (red dots) and ACN:MeOH:H_2O (blue dots). 20
Figure 3.2.5 - Spectra from the different solvent extraction assay. (A) Spectra resultant from the application of Atmospheric correction to eliminate potential CO_2 and H_2O noise, Baseline Correction, Rubber band method, and a posterior Maximum Normalization executed on The Unscrambler® X software. (C) Spectra resultant from the application of Atmospheric correction, Baseline Correction and a posterior Unit Vector Normalization. B and D are the corresponding PCA's for A and C, and here disposed for a better observation of the impact of each normalization done in the data. 21
Figure 3.2.6 - Savitzky-Golay application. As observed multiple derivatives (yellow lines) are being applied tangent to raw data (in blue) and the segment of that derivate (red line) is being used to fit the new data sub-set that will be the new data (blue circles). 21
Figure 3.2.7 - Spectra from the different solvent extraction assay. (A)Spectra data relative to extraction assay, which suffered Atmospheric correction on OPUS software followed by Savitzky-Golay Transformation with a

derivative Order of 2, Polynomial Order of 2, with 15 smoothing points. (C) Is the same spectra present in (A) but with a wavelength cut applied, 800-1800-2800-3600 cm^{-1}. 22
Figure 3.3.1 - Normal workflow followed during MS analysis. Data from UPLC-ESI-HRMS/MS are converted to a compatible format and uploaded into XCMS online servers, where group comparison are made. XCMS treats data in an automated way, such as, chromatographic peak detection and retention time alignment, as well mass spectra matching [111]. From there various plots are generated, like, boxplot for the *m/z* intensities of each group and cloud plot for significant features detected. XCMS uses METLIN open-source database to identify features into predictive metabolites [112–115]..... 23
Figure 3.3.2 - Example of an Extracted Ion Chromatograms (XIC). Like Total Ion Chromatograms (TIC) where all intensities of each spectre scan are summed up and reflected on the graphic as a total intensity at a given retention time, XIC graphic give us, for a given retention time, the most intense observation of that spectral scan and returns it as total intensity. In the Y axis is the total ion intensity and X axis the retention time (in minutes). This XIC is the representation of the signal acquired from the extraction method assay [117]. Platform used for visualization: MzMine 2.53 [115]. 24
Figure 3.3.3 - Examples of data collected from XCMS online. (A) Cloud Plot generated from pairwise comparison of 2 distinct groups in the same LC-MS acquisition mode, in this case is group A vs group B (group explanation in chapter 3.4) in the same experimental setup. From the pairwise job both, a chromatogram alignment (B) as a dual boxplot comparison for a given *m/z* (treated as a feature) is retrieved to compare each group. Finally, a multimodal job (combines identified data by sample only, regardless of chromatographic or spectrometric acquisition conditions is executed, and the table retrieved is like shown in (C) there all features detected in the pairwise comparison are compared against the METLIN [114] human metabolome database for metabolite identification. 25
Figure 3.4.1-Representative scheme of patient and sample logistics. 6 patients were chosen in accord with variables shown in this chapter, and each sample was processed with the two extraction methods, separately. Each extracted sample was divided into three aliquots that were stored in -80 ºC till FTIR and MS analysis, totalizing 36 samples. Samples were further identified as GROUP A, B, C and D according to extraction submitted and according to the patient's outcome later explained. 25
Figure 3.4.2 - Representative scheme of patient and sample logistics. 24 patients were chosen in accord with variables shown in this chapter. Samples were processed using the ACN:MeOH:H_2O method. The total 24 resultant aliquots were stored at -80ºC. 26
Figure 4.1.1.1 – (A) Spectral data retrieved from FTIR analysis, corresponding to Raw data (no pre-processing). (C) Spectral data when performed Atmospheric Correction. (B) Scores diagram from data (A) with Hotelling's T2 ellipse at 5% significance for PC1 (90% variance) versus PC2 (5% variance). (D) Scores diagram from data (C) with Hotelling's T2 ellipse at 5% significance for PC1 (50% variance) versus PC3 (10% variance). Program used, The Unscrambler X®..... 28
Figure 4.1.1.2 - Loading plot for PC3 from figure 4.1.1.1(D). Loading scores are always limited between 1 and -1, which are expressed on the Y axis of the plot. On the X axis, are the corresponded variables that the principal components use, which are the wavenumbers (from 400 to 4000) used for FTIR analysis. 28
Figure 4.1.1.3 – Loading scores from PC3 enlightened in a red circle, majorly responsible for the separation of the two clusters on the PCA (right) from figure 4.1.1.1(C). 29
Figure 4.1.1.4 - (A) Spectra resultant from Atmospheric and Baseline correction, and Unit Vector Normalization. (B) PCA (PC1 63% variance against PC2 17% variance) correspondent to spectra (A). (C) Spectra resultant from Atmospheric correction and Savitzky-Golay filter with a 2nd order polynomial and a 15-points window. (D) PCA (PC1 with 68% variance against PC2 with 4% variance) resultant from spectral data (B). 30
Figure 4.1.1.5 - Hierarchical Cluster Analysis for serum pre-processed with atmospheric and baseline correction with unit vector normalization. HCA complete linkage method and cosine correlation distance metric was used. Medians for replicas were used for this analysis. 8/36 samples were misclassified, corresponding to samples 31, 32 and 34 (corresponding to extraction samples from patient 6 submitted to methanol extraction); sample 24 (sample from patient 4 from ACN:MeOH:H_2O); samples from patient 5 (sample number 28 and 29 from ACN:MeOH:H_2O, and 25 and 26 from methanol extraction). Program used, Orange 3.32. 31
Figure 4.1.1.6 - Predictive model based on the 400 to 4000 cm^{-1} wavenumbers for patient outcome. 144 samples total ran through the Stochastic Gradient Descent model, ran on default parameters on Orange 3.32. 32

Figure 4.1.1.7 – Workflow for prediction using various machine learning algorithms. HCA analysis was conducted for all replicates (144 samples), and for median values (36 samples). This workflow ran three times, corresponded to the three data pre-processing techniques chosen. Consider that analysis took over one hour for each pre-processing due training and test methods adopted. 33
Figure 4.1.2.1 – (A) Spectra from data pre-processed with atmospheric and baseline correction from all patients of second study assay. (B) PCA from spectra mentioned, PC1 with 78% variance against PC2 with 17% variance. 33
Figure 4.1.2.2 – Explained variance for PCA (A) before patient 40 second replicate removal and (B) after removal. (blue line: calibration; red line: validation). 34
Figure 4.1.2.3 – PCA score (PC2 13% variance against PC4 3% variance) from dataset pr-processed with atmospheric and baseline correction followed by maximum normalization, which divides each absorbance from a sample by its maximum value. All samples with all replicates from all groups were utilised apart from sample from patient 40 replicate number 2. 34
Figure 4.1.2.4 – Loading Score from PC2 from figure 4.1.2.3. As observed this principal component has a large loading weight >0.4 from wavenumber ~3346 to 3226 cm^{-1} and another peak at ~1655 cm^{-1}. From the 800 cm^{-1} below the loading presents heavy negative values despite of the PCA report saying that this band was not impacting the Score..... 35
Figure 4.1.2.5 – (A) PCA (Pc2 against PC7) between samples from patients who did not go through ventilation (Group A) and those who went through ventilation (Group B), both survived. (B) Principal component 2 loading score, the main PC that separates group A and B. It can be observed the main wavenumbers that influence this PC. 36
Figure 4.1.2.6 - A) PCA (PC2, 17% variance against PC4, 3% variance) between samples from patients who did not go through ventilation and survived (Group A) and those who went through ventilation and died (Group C). (B) Principal component 2 loading score, the main PC that separates group A and C. It can be observed the main wavenumbers that influence this PC. 37
Figure 4.1.2.7 - A) PCA (PC4 against PC6) between samples from patients who went through ventilation and survived (Group B) and those who went through ventilation and died (Group C). (B) Principal component 6 loading score, the main PC that separates group B and C. It can be observed the main wavenumbers that influence this PC. 38
Figure 4.1.2.8 – Workflow performed on the three analyses from Chapter 4.1.2. 39
Figure 4.2.1.1 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.1. (α-Cehc is a synonym of α-carboxyethylhydroxychroman). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA resultant from allocating the m/z intensity of each metabolite of each group (A: Blue; B: Red) in The Unscrambler® X, is possible to see a fine discernment between group A and group B, being B better clustered in relation to A. 40
Figure 4.2.1.2 - Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.2. Program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA resultant from allocating the m/z intensity of each metabolite of each group (C: Blue; D: Red) in The Unscrambler® X, is possible to see a fine discernment between group C and group D, being in this case, D better clustered compared to D. 41
Figure 4.2.1.3 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.3. Each metabolite is represented in both group A (left) and C (right). Y axis is expressed to the power of 0.3. On the X axis each metabolite is represented from left to right by the order indicated on the plot, program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (Abbreviation: 4-hydroxy-2-NGC: 4-hydroxy-2-nonenal-glutathione conjugate). (B) PCA (PC 1 against PC 5) resultant from allocating the m/z intensity of each metabolite of each group (A: Blue; C: Red) in The Unscrambler® X. Although clustering could not be well achieved for both deceased and survivor patient samples, PC1 against PC5 could separate both groups. 42
Figure 4.2.1.4 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.4. Program used IBM SPSS Statistics. (Abbreviations: Zymosterol: 4α-OCH3-4β-CH-5α-zymosterol; Decaprenylbenzoate: 3,4-dihydroxy-5-all-trans-decaprenylbenzoate). (B) PCA (PC3 against PC4) resultant from allocating the m/z intensity of each metabolite of each group (B: Blue; D: Red) in The Unscrambler®. 43

Figure 4.2.2.1 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.1. Y axis is expressed to the power of 0.5. Program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC4) resultant from allocating the m/z intensity of each metabolite of each group (A: Alive Non-ventilated patients; B: Alive Ventilated Patients) in The Unscrambler®. 44
Figure 4.2.2.2 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between non-ventilated and ventilated alive patients. 45
Figure 4.2.2.3 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.2. Program used IBM SPSS Statistics. (Abbreviations: Benzoquinol: 6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol; FMN: Flavin Mononucleotide). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC4) from resultant from allocating the m/z intensity of each metabolite of each group (A: Alive Non-ventilated patients; C: Ventilated deceased patients) in The Unscrambler®. 46
Figure 4.2.2.4 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between non-ventilated alive and ventilated deceased patients samples. 47
Figure 4.2.2.5 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.3. Intensity for the given metabolite (y-axis) is expressed to the power of 0.5. Program used IBM SPSS Statistics. (Abbreviations: Butanoate: 4-(3-pyridyl)-butanoate). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC3) resultant from allocating the m/z intensity of each metabolite of each group (B: Alive Ventilated patients' samples; C: Deceased Ventilated patients' samples) in The Unscrambler®. 48
Figure 4.2.2.6 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between ventilated alive and deceased patients samples. 49
Figure 4.2.2.7 – workflow for learning and prediction analysis for chapter 4.2.2. 50
Figure 1 – 1st assay EIC's of (A) α-tocopherol; (B) (R)-lactate; (C) Cytidine; (D) 4α-hydroxymethyl-4β-methyl-5α-zymosterol; (E) α-carboxyethylhydroxychroman; (F) (R)-3-hydroxybutanoate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group B (red). 62
Figure 3- 1st assay EIC's of (A) D-sorbitol; (B) 4-(3-pyridyl)-butanoate; (C) 3'-dephospho-CoA; (D) (R)-3-hydroxybutanoate; (E) (R)-lactate; (F) stearate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group C (Black) and Group D (red). 63
Figure 4 – 1st assay EIC's of (A) α-D-glucose; (B) α-L-iduronate; (C) D-sorbitol; (D) biliverdin-IX-α; (E) palmitate; (F) α-linolenate; (G) stearate; (H) 4-hydroxy-2-nonenal-glutathione conjugate; (I) 6-pyruvoyl tetrahydropterin; (J) (4Z,7Z,10Z,13Z,16Z,19Z)-docosahexaenoate; (K) 4-(3-pyridyl)-butanoate; (L) N-acetyl-β-neuraminate; (M) 2-trans-hexadecenal. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group C (red). 65
Figure 5 - 1st assay EIC's of (A) hypoxanthine; (B) biliverdin-IX-α; (C) D-sorbitol; (D) N-formylkynurenine; (E) 3,4-dihydroxy-5-all-trans-decaprenylbenzoate; (F) sulfate; (G) α-L-iduronate; (H) 4-(3-pyridyl)-butanoate; (I) palmitate; (J) N-acetyl-β-neuraminate; (K) 4α-hydroxymethyl-4β-methyl-5α-zymosterol. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group B (Black) and Group D (red). 66
Figure 6 - 2nd assay EIC's of (A) L-ascorbate; (B) bilirubin; (C) thiamin; (D) N-acetyl-β-neuraminate; (E) 5-methoxytryptophol; (F) calcitriol; (G) aldehydo-D-glucuronate; (H) L-gulonate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group B (red) 67
Figure 7 - 2nd assay EIC's of: (A) Octanoate; (B) Thiamin; (C) Sucrose; (D) α-L-iduronate; (E) 4-acetamidobutanoate;(F) 6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol; (G) FMN; (H) 3'-dephospho-CoA; (I) α,α-trehalose. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group C (red). 69

Figure 9 - 2nd assay EIC's of: (A) 4-(3-pyridyl)-butanoate; (B) octanoate; (C) phenylacetate; (D) 5,6-dihydrouracil; (E) stearate; (F) linoleate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group B (Black) and Group C (red)..... 70

List of Tables

Table 2.1.1 - Non-Structural proteins coded by SARS-CoV-2 Genome..... 6
Table 2.3.1 – Differences between the two major metabolic analytical platforms. Adapted from Attard *et al* [93]. 10
Table 2.3.2 – Main differences between Untargeted and Targeted Metabolomic Analyses. Adapted from Schrimpe *et al* [97]..... 11
Table 2.3.3 - Overview of results and models from published studies using metabolomics to characterize COVID-19 patients. 13
Table 2.4.1 – Main features of IR spectroscopy. Adapted from Wade *et al* [107]..... 16
Table 2.4.2 - Vibrational frequencies in FTIR spectroscopy for serum. Adapted from Wood *et al* [108], Movasaghi *et al* [109], Cakmak *et al* [110]. 16
Table 3.3.1 - Parameters used in XCMS Online analysis for patients' samples untargeted metabolomics data process. 24
Table 3.4.1 - Patients' demographics. Due to sample size, each variable is expressed for each patient to facilitate assessment. (Abbreviations: BMI: Body Mass Index; ECMO: Extracorporeal Membrane Oxygenation; ICU: Intensive Care Unit; PCR: Polymerase Chain Reaction) 26
Table 3.4.2 - Patients' demographics. (Abbreviatures: BMI: Body Mass Index; IQBV- Inter-Quartile Boundary Values [Percentile 25; Percentile 75]; COPD- Chronic Obstructive Pulmonary Disease; OSA- Obstructive Sleep Apnoea; ECMO: Extracorporeal Membrane Oxygenation). 27
Table 4.1.1.1 - Samples identified as outliers in replicates of FTIR serum spectra from the 6 patients samples, extracted with methanol and ACN:MeOH:H2O extraction protocols, and submitted to the three best considered pre-processing corrections. 30
Table 4.1.1.2 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for target (death / survival): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall). 32
Table 4.1.1.3 – Confusion matrix for patient outcome, from SGD regression model. Percentages relative to proportion of predicted samples. 32
Figure 4.1.2.1 – (A) Spectra from data pre-processed with atmospheric and baseline correction from all patients of second study assay. (B) PCA from spectra mentioned, PC1 with 78% variance against PC2 with 17% variance. 33
Table 4.1.2.1 – Outlier identification. From dataset pre-processed with atmospheric and baseline correction followed by maximum normalization. 35
Table 4.1.2.2 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for target (Ventilation): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall). 36
Table 4.1.2.3 - Confusion matrix for patient condition: Ventilation status, from Neural Network prediction model. Percentages relative to proportion of predicted samples. 36
Table 4.1.2.4 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for outcome, between samples from non-ventilated alive patients (group A) and ventilated deceased patient samples (group C): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall). 37
Table 4.1.2.5 - Confusion matrix for patient outcome: death, from Neural Network prediction model. Percentages relative to proportion of predicted samples. 38
Table 4.1.2.6 - Test and scores results for all learning methods. Sampling type: Leave One Out (LOO); Evaluation results for outcome, between samples from ventilated alive patients (group B) and ventilated deceased patient samples (group C): average over class. Models were run on default setup offered by Orange 3.32 software.

(Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall). 38
Table 4.1.2.7 - Confusion matrix for patient outcome and condition: death and ventilation status, from kNN prediction model. Percentages relative to proportion of predicted samples. 39
Table 4.2.1.1 - Metabolites Results from subgroup A vs B. (Abbreviations: R.T: retention time; F.C.: fold-change). Dysregulation in this context means how much higher the intensities of the metabolite of group B is compared to group A (e.g., α-tocopherol has a 3.99 fold-change with dysregulation "Down", this mean that group A has an intensity 3.99 times higher than B for the given molecule, and so, B is down-regulated relative to A). 40
Table 4.2.1.2 - Metabolites Results from subgroup C vs D. (Abbreviations: R.T: retention time; F.C.: fold-change). 41
Table 4.2.1.3 - Metabolites Results from subgroup A vs C. (Abbreviations: R.T: retention time; F.C: fold-change). 42
Table 4.2.1.4 - Predictive Metabolites Results from subgroup B vs D. (Abbreviations: R.T: retention time; F.C.: fold-change). 43
Table 4.2.2.1 - Metabolites Results from the comparison between samples from COVID-19 patients who didn't need ventilation and who went through ventilation (group A versus B, respectively). (Abbreviations: R.T: retention time; F.C.: fold-change). 44
Table 4.2.2.2 - Metabolites Results from group A vs C (nonventilated alive patients against ventilated deceased patients). (Abbreviations: R.T: retention time; F.C.: fold-change; FMN: Flavin Mononucleotide). 46
Table 4.2.2.3 - Metabolites Results from subgroup B vs C. (Abbreviations: R.T: retention time; F.C.: fold-change). 48
Table 4.2.2.4 – Machine learning results from samples retrieved from mass spectrometer data from the second assay patients' samples. Models computed to determine in each data set the patient submission to ventilation (group A vs B); outcome between non-ventilated and ventilated deceased patients (group A vs C); outcome between ventilated patients (Group B vs C). Sampling type: Leave One Out (LOO); Evaluation results for target: average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall). 50
Table 4.2.2.5 – Confusion matrix retrieved from neural network. Ventilated discharged patients versus ventilated deceased patients (group B vs C). Percentages relative to proportion of predicted samples..... 50

List of Abbreviations

+ssRNA	Positive Single-Strand RNA
3CLpro	3-Chymotrypsin-like protease
A.M.U	Atomic Mass Unit
ACE	Angiotensin Converting Enzyme
ACE2	Angiotensin-Converting Enzyme 2
AdaBoost	Adaptative Boosting
APC	Antigen-Presenting Cells
ARDS	Acute Respiratory Distress Syndrome
ATM	Atmospheric Correction
AUC	Area Under the Curve
BC	Baseline Correction
BMI	Body Mass Index
CA	Classification Accuracy
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 2019
CoVs	Coronaviruses
CT	Computerized Tomography
DAMP	Damage-Associated Molecular Pattern
DMV	Double Membrane Vesicle
DNA	Deoxyribonucleic Acid
DPD	Dihydropyrimidine Dehydrogenase
ECMO	Extracorporeal Membrane Oxygenation
EIC	Extracted Ion Chromatograms
EM	Electron Microscopy
ERGIC	Endoplasmic Reticulum Golgi Intermediate Compartment
ESI	Electrospray ionization
F.C.	Fold Change
F1	F1-score
FTIR	Fourier transform infrared
GABA	Gamma Aminobutyric Acid
HCA	Hierarchical Cluster Analysis
HILIC	Hydrophilic interaction liquid chromatography
HMDB	Human Metabolome Database
HMR	High-Resolution Mass Spectrometry
ICTV	International Committee on Taxonomy of Viruses
ICU	Intensive Care Unit
IQBV	Inter-Quartile Boundary Values
IR	Infra-Red
IVM	Invasive Mechanical Ventilation
kNN	k-nearest neighbours
LC-MS	Liquid Chromatography coupled to Mass Spectrometry
LOO	Leave One Out
m/z	Mass to charge
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
MHC	Major Histocompatibility Complex
mRNA	Messager-RNA
MS	Mass Spectrometry
MSSA	Methicillin-Susceptible Staphylococcus Aureus
NK cell	Natural Killer Cell
NMR	Nuclear Magnetic Resonance
NSP	Non-Structural Protein

ORFs	Open Reading Frames
OSA	Obstructive Sleep Apnoea
PAMP	Pathogen-Associated Molecular Pattern
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PLpro	Papain-like protease
PREMO	*Predictive Models of COVID-19 Outcomes for Higher Risk Patients towards a Precision Medicine*
QqTOF	quadrupole-quadrupole time-of-flight
R.T.	Retention Time
RAS	Renin-Angiotensin System
RdRp	RNA-dependent RNA polymerase
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
ROC	Receiver Operating Curve
RP	Reverse Phase
RTC	Replication-Transcription Complex
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SG	Savitzky-Golay
SGD	Stochastic gradient descent
sgRNA	sub-genomic RNA
SVM	Support-vector machine
TCA	Tricarboxylic Acid Cycle
TMPRSS2	Transmembrane Serine Protease 2
tRNA	Transfer RNA
UHPLC-MS	Ultra-High-Performance Liquid Chromatography coupled to Mass Spectrometry
UVN	Unit Vector Normalization
WHO	World Health Organization

[Blank Page]

Chapter 1

1.1. Introduction

Coronaviruses are a diverse group of viruses capable of triggering mild to severe respiratory infections on the Human species. In 2002 and 2012, two zoonotic viruses, SARS-CoV (*Severe Acute Respiratory Syndrome Coronavirus*) and MERS-CoV (*Middle East Respiratory Syndrome Coronavirus*) emerged, causing serious damage on the respiratory tract on the Human species [2].

In late 2019, a new coronavirus variant arose in many hospital units in Wuhan, China, causing an endemic viral outbreak. This new variant, highly transmissible, surpassed both SARS and MERS in terms of the number of infected individuals. Similarly to the former, the new virus symptoms included pneumonia, fever, cough and chest discomfort, and in severe cases, dyspnea and bilateral lung infiltration [3,4].

Through RNA metagenomic sequencing of the recovered virus from bronchoalveolar lavage fluid from patients with severe pneumonia, it was possible to identify that scientists were dealing with a new and unidentified coronavirus [5].

Meanwhile, the incidence in Wuhan spread to other Chinese provinces, thanks to the heavy population flow due to the New Year Chinese Festival. On January 30, WHO declared the novel coronavirus outbreak a public health emergency of international concern, naming it 2019-nCoV [6]. In february 2020, WHO (World Health Organization) officially declared the outbreak of the disease as COVID-19, while the ICTV (International Committee on Taxonomy of Viruses) recognizes this new virus as a SARS-CoV genetic variant related to species *Severe Acute Respiratory Syndrome-Coronavirus*, and designates the new species as SARS-CoV-2 [7,8].

In March 2020, the WHO classified the disease as a pandemic [9]. Despite the genetic evidence pointing out that SARS-CoV-2 is a virus present in nature and of zoonotic origin, there is still no consensus on when and how the virus entered the human species. Some experts pointed to its presence in the seafood market in Wuhan, however, with the development of the pandemic, fewer and fewer cases were linked to the Wuhan market, pointing to origins beyond this, with human transmission currently being the largest vector of spread of the disease [5,10].

1.2. Objectives

Due to the relevance of the COVID pandemic, this thesis aims to develop a predictive model for the diagnosis of the outcome of this disease, as well as for the prognosis in critically ill patients, based on a metabolomic analysis of their serum. This is an observational and prospective study, without interference in the assessment, diagnosis, and therapy of patients. This work is part of a larger project, already approved by the Ethics Committee of the Hospital (Centro Hospitalar Universitário Lisboa Central), under all legal and ethical considerations. Samples will be analyzed using Fourier-Transform Infrared spectroscopy (FTIR) and UHPLC-MS/MS untargeted metabolomics workflows.

Samples will be analyzed by a microplate FTIR reader for high-throughput screening IR spectroscopy used for measurements in the mid infrared range, and each serum sample will be divided into aliquots for Reverse Phase (RP) / UHPLC-MS / MS analysis, ESI operated in both modes (±), and for Hydrophilic interaction liquid chromatography (HILIC) / UHPLC-MS / MS analysis, ESI (±). A high-resolution quadruple time of flight (QqTOF Impact II) mass spectrometer will be used.

Chapter 2 - Literature Review

2.1. SARS-CoV-2 Taxonomy, Characterization, and Replication

Coronaviruses (CoVs) are a family of non-segmented, encapsulated, nucleocapsid, positive-strand RNA (+ssRNA) viruses that infect vertebrates [8]. They are classified under the kingdom *Riboviria*, order *Nidovirales*, family *Coronaviridae*, subfamily *Orthocoronavirinae*, and are divided into four genera, *Alphacoronavirus* (αCoV), *Betacoronavirus* (βCoV), *Deltacoronavirus* (δCoV), and *Gammacoronavirus* (γCoV) (figure 2.1.1) [8,11,12].

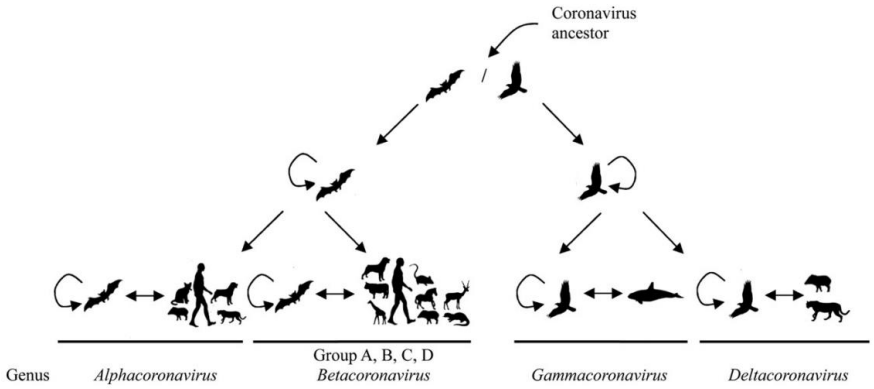


Figure 2.1.1 - Evolution model of the *CoV* species. Adapted from Woo *et al*. [13].

Phylogenetic analysis of SARS-CoV-2 showed that this species shows homology with SARS-CoV (79%) and with the MERS-CoV species (50%), as well as a correspondence between bat-SL-CoVZC45 coronaviruses (sequence identity 87.99%) and another SARS-like beta coronavirus of bat origin, bat-SL-CoVZXC21 (87.23%) as described in Lu *et al*. [14]. Thus, this new species is classified in the genus *Betacoronavirus*, subgenus *Sarbecovirus*, (figure 2.1.2).

Category	Coronaviruses	Humans	Divergence
Realm	*Riboviria*		
Order	*Nidovirales*	Primates	
Suborder	*Cornidovirineae*		
Family	*Coronaviridae*	Hominidae	
Subfamily	*Orthocoronavirinae*	Homininae	
Genus	*Betacoronavirus*	*Homo*	
Subgenus	*Sarbecovirus*		
Species	*Severe acute respiratory syndrome-related coronavirus*	*Homo sapiens*	
Individuum	SARS-CoVUrbani, SARS-CoVGZ-02, Bat SARS CoVRf1/2004, Civet SARS CoVSZ3/2003, SARS-CoVPC4-227, SARSr-CoVBtKY72, SARS-CoV-2 Wuhan-Hu-1, SARSr-CoVRatG13, and so on.	Dmitri Ivanovsky, Martinus Beijerinck, Friedrich Loeffler, Barbara McClintock, Marie Curie, Albert Einstein, Rosalind Franklin, Hideki Yukawa, and so on.	

Figure 2.1.2 - Taxonomy of coronaviruses compared to the human species. Adapted from Gorbalenya *et al*. [8].

Electron Microscopy (EM) of SARS-CoV-2 particles revealed their spherical shape, with diameters varying between 60 and 140 nm, and an outer capsular surface full of distinct crownlike spikes (Spikes), 9-12 nm in length, giving the appearance of a solar corona, hence its name [15]. Schematic representation of the main functional SARS-CoV's subunits are shown on figure 2.1.3.

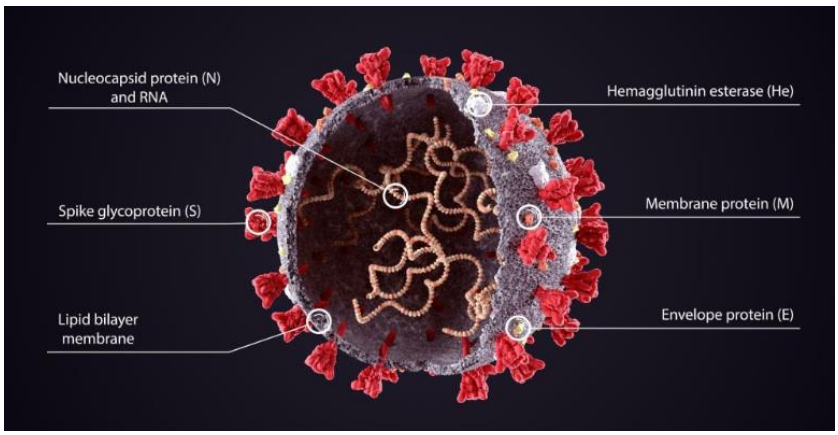


Figure 2.1.3 - Schematic representation of the structure of SARS-CoV-2. Figure 2.1.3 shows an unreported structure, Hemagglutinin-esterase (He) is a glycoprotein existing in some viruses in the envelope zone, which helps in the cell invasion mechanism. However, this molecule is not always identified in the structural composition of the SARS-CoV-2 species. (Image: Orpheus FX/Shutterstock)

CoV genomes (27-32 kb) encode proteins such as: Nucleocapsid (N) that complexes with the virion genome, and an outermost capsular structure formed by three structural proteins: membrane protein (M), spike protein (S), and the envelope protein (E).

The SARS-CoV-2 genome, measuring 29.9 kb (29903 nucleotides encoding a total of 9860 amino acids [16]), is "packed" by proteins, the nucleocapsids (N), forming a ribonucleoprotein (RNP) complex [17], which in turn is included in an envelope membrane with lipids and viral proteins, the protein S (surface or Spike), M (membrane) and E (envelope) [18]. Its genome encodes the structural genes (N, M, S and E) and sixteen non-structural proteins (nsp 1-16) genes [19].

The SARS-Cov-2 genome includes 6 ORFs (Open Reading Frames), and at the 5' end of its genome there are two large ORFs, ORF1ab, occupying about two thirds of the genome. The translation of ORF1ab is mediated by a -1 frameshift (also known as Programmed -1 ribosomal frameshifting, or translation matrix shift) that "delays" the reading frame by one nucleotide and allows the ribosomal complex to continue past the STOP codon of the ORF1ab. ORF1a, allowing from the same mRNA strand (ORF1) to obtain 2 distinct reading frames, designated in the case of SARS-Cov-2 as pp1a and pp1b (polyprotein 1a, and polyprotein 1b respectively) which are subsequently cleaved into 16 non-structural proteins (nsp1-nsp16). A representation of the Frameshifting phenomenon can be seen in figure 2.1.4 [20,21].

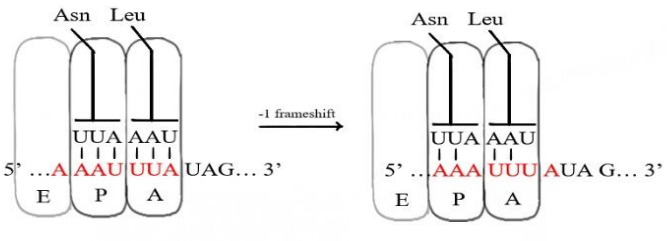


Figure 2.1.4 – Frameshifting process. The three pill-like shaped figure represents the ribosomal complex, while the single stranded RNA template is represented by the line of letters, some of which are red coloured. The tRNA here transports 2 distinct amino acids, asparagine (Asn) and leucine (Leu).

In its genomic composition, SARS-CoV-2 encodes 16 non-structural proteins (nsp's), 4 structural, and 6 accessory proteins, as can be seen in figure 2.1.5. All 16 nsp's, described in table 2.1.1, are encoded in 5' end of the genome, and are related to replication, viral transcription and evasion of the host's immune system, their functions as well as those of other ORFS have been studied for a better understanding of the structure, propagation, and viral pathogenicity patent in this species [22–25].

Proteins that contribute to the virions' structure (structural proteins) as well as accessory proteins (also known as 3a, 6, 7a, 7b, 8 and 9b), with functions yet to be determined, are encoded by the rest of the viral genome at the 3' end. It represents 30% of the virus genome compared to ORF1ab which represents 70%. There are also other ORFs, such as ORF10, which is annotated in databases such as GenBank, and which is computationally predicted, but which, however, in terms of protein expression is not denoted, representing trace sequences in sequencing techniques by genome amplification [17]. In spite of the fact that this findings reveal several functions and characterization of the new SARS-CoV-2 genome, there is still evidence of the importance of more detailed studies in the sequencing and validation of the expression of sgRNAs (sub genomic RNA) of the SARS-CoV-2 species [18,20,26,27].

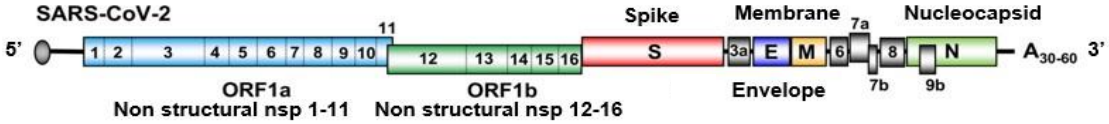


Figure 2.1.5 - SARS-CoV-2 genomic structure. composed of ssRNA+ with a cap (grey circle) at the 5' end and a poly-A tail at the 3' end. It encodes 16 nsp's (ORF1a → nsp1-11 and ORF1B → nsp12-16). Structural proteins are identified as S, E, M and N and accessory proteins are numbered and identified by gray rectangles. Adapted from Brant *et al*. [18]

Being a positive sense RNA virus (+ ssRNA), this species can translate the two polyproteins encoded in ORF1ab in the cytoplasm of the susceptible cell. Cleavage of this protein complex is then carried out by two self-activating proteases (Papain-like protease or PLpro, also known as nsp3 and 3-Chymotrypsin-like protease 3CLpro or nsp5), giving rise to 16 nsp's. The RNA (+) strand is initially duplicated to an RNA (-) strand, which is then utilized for either replication to another RNA (+) strand for fresh virion assembly or transcription of sub-genomic mRNA that may be translated into a variety of viral proteins.

These sgRNAs have a feature in common, they share a cap- structure at the 5' end (5'-cap, ~70 nt's) and a poly-AAA tail at the 3' end (multiple adenosine nucleotides, with 30-45 nt's in length), for viral genome stability and prevention of digestion by cellular exoribonucleases [18,26].

The first stage of cellular infection begins with the binding of the virus to the host cell, initiated by the interaction of the S protein with the host receptor, the angiotensin-converting enzyme ACE2 (angiotensin-converting enzyme 2) [22,28] (figure 2.1.6). Studies indicate the existence of other factors promoting virion-cell adhesion as reported in [29–32] as the initial adhesion of the virion to the cell is carried out by the cell surface polysaccharide, heparin sulfate, present in the cells [33]. Proteases such as TMPRSS2 (transmembrane enzymes) and Cathepsin L cleave the S protein in the S1 subunit responsible for anchoring (also known as the region-binding domain, RBD, which has been targeted for therapy with neutralizing antibodies [34]), and S2 subunit, responsible for virion-cell fusion, in attempt to activate SARS-CoV-2 entry by endocytosis and membrane fusion [35] (figure 2.1.7, 1).

Both ACE2 and TMPRSS2 are expressed in various cell types, particularly in the lungs (type II pneumocytes), gut epithelium (enterocytes), corneal endothelial cells, bladder epithelium, and nasal epithelium (nasal goblet cells), allowing SARS-CoV-2 to reach various vital organs [35–39].

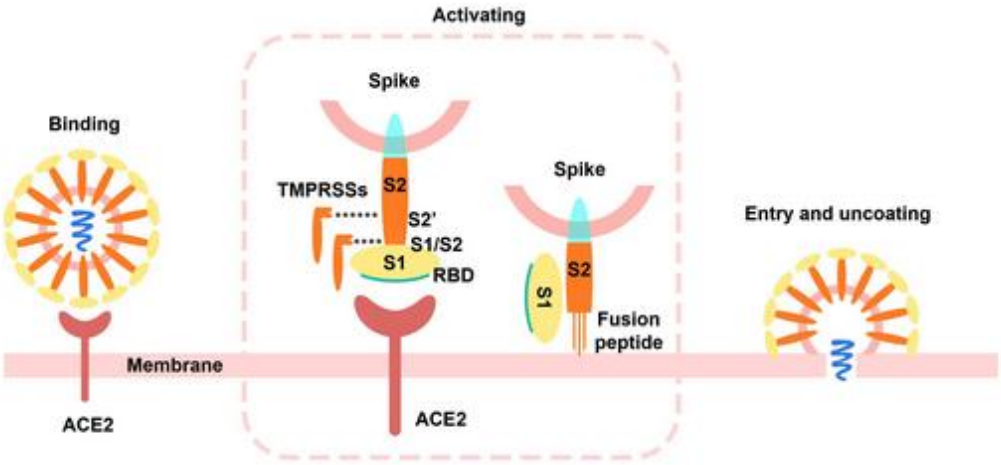


Figure 2.1.6 - Spike protein role in host cell invasion (image: AIVD Biotech Inc.).

As an RNA virus, it replicates exclusively in the cytoplasm of the host cell, where the positive RNA genome is first disassociated from the nucleocapsids (Protein N) (figure 2.1.7, 2). From here the genomic RNA is used directly by the host cell's ribosomal machinery to carry out the translation of ORF1a (giving rise to pp1a) and by the frameshifting mechanism to read ORF1ab (giving rise to pp1ab) autocatalytically processed by viral proteases located in nsp3 (PLpro) and nsp5 (3CLpro or Mpro, from main protease) in the 16 nsp's (figure. 2.1.7 ,3, 4), which establish a replication-transcription complex (RTC) in this step of the infection the nsp1 neutralizes host translation as well as cellular mRNA, while nsp2-16 sets the RTC. RTC is a protein complex that includes key proteins in virus replication, such as RNA-dependent RNA polymerase (RdRp, or replicase), proteins with helicase, primase, endo- and exoribonuclease activity, and NSP16. The RTC complex is located in a cell membrane remodeled by nsp's, such as nsp6, which form continuous replication organelles with the endoplasmic reticulum, providing an optimized environment for viral RNA replication [40,41] and protection against detection by sensors. cytosolic innate immune cells [42] (figure. 2.1.7 ,4b).

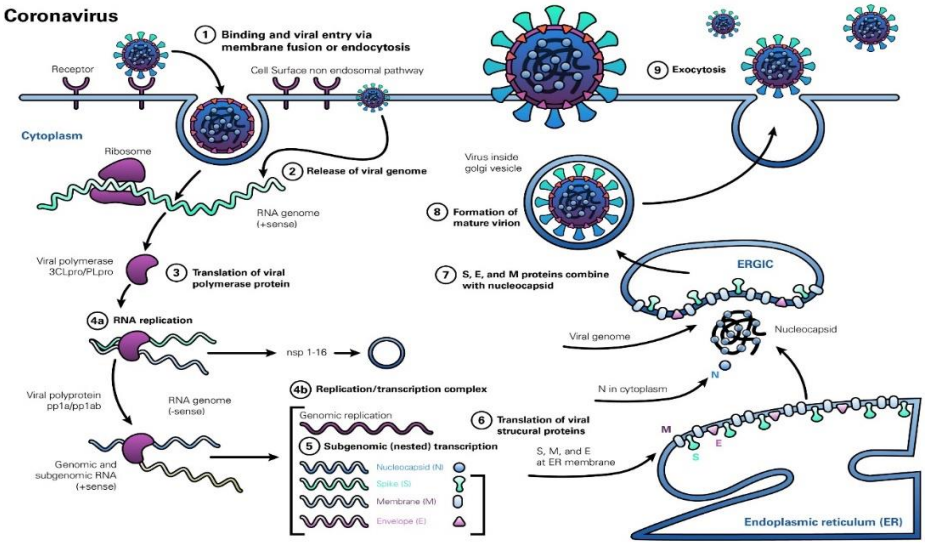


Figure 2.1.7 – SARS-CoV-2 life cycle. Adapted from Burmer *et al.* [43]

Table 2.1.1 - Non-Structural proteins coded by SARS-CoV-2 Genome.

Protein	Function	Ref.
NSP1	Involved in restraining the innate antiviral response of the host and in suppressing the induction of apoptosis during early infection to promote viral growth.	[44]
NSP2	Involved in the disruption of host intracellular signalling during SARS-CoV infections.	[45]
NSP3	Supposedly involved in the translation of mRNA transcripts and suppression of host protein synthesis.	[46]
NSP4	Essential role in the replication and assembly of replicative structures.	[46,47]
NSP5	Protease activity.	[48]
NSP6	Synthesis of autophagosomes from the endoplasmic reticulum and involved in the process of autophagy (DMV's - Double Membrane Vesicle).	[46,49]
NSP7	RNA dependent RNA polymerase activity.	[50]
NSP8	Primase activity	[51]
NSP9	Complexes with NSP8, involved in RNA replication and virulence.	[52]
NSP10	It is a cofactor for the activity of NSP16 2′O-methyltransferase and N7-guanine methyltransferase/exoribonuclease of NSP14.	[53]
NSP11	Essential for replication, a more precise role is yet to be determined.	[54]
NSP12	RNA polymerase/replicase activity.	[11,55]
NSP13	Helicase and RNA NTPase activity.	[56,57]
NSP14	Methyltransferase and exoribonuclease activity.	[53,58]
NSP15	Uridylate-specific endoribonuclease activity.	[59]
NSP16	Methyltransferase that mediates cap-2'-O-ribose methylation of mRNA to the 5'-cap structure of viral mRNAs.	[53,60]

RNA replication begins with the complete synthesis of genomic copies. These serve as a template to produce new positive strand copies, used to produce more nsp's or to be inserted into a new viral structure to give rise to a mature virion. At the same time, segmented transcription of the genome occurs at the level of the RTC-DMV complex, which generates negative-stranded subgenomic RNA, which serve as a template for the synthesis of positive-stranded subgenomic RNA (sgRNA+) [61]. These RNAs are translated to synthesize structural and accessory proteins (figure. 2.1.7, 5, 6).

The synthesized viral RNA molecules escape through transmembrane pores existing in the DMV's [62], and are mobilized to the site of translation and assembly of the virion that occurs in the endoplasmic reticulum and intermediate compartment of the Golgi complex – endoplasmic reticulum (ERGIC), respectively [63] (figure 2.1.7, 7). Here, newly synthesized viral genomes coated with the N protein germinate in the lumen of the ERGIC to form enveloped particles containing the M, E and S proteins. Finally, this structure is complexed with the viral gRNA to form the final mature virion that will be transferred to the host cell surface to be secreted (figure 2.1.7, 9).

2.2. SARS-CoV-2 Transmission and Pathogenicity

The primary route of human-to-human transmission of COVID-19 disease is through droplets that pass through the airway mucosa and come into direct or indirect contact with another individual or fomite, respectively [64,65]. Frequent coughing as well as sneezing exhibited by COVID-19 patients generate clouds of thousands of droplets per cubic centimeter [66], Thus, it is imperative to understand the dynamics of transmission as well as the different routes of passage/intrusion of the virus through the physiological barriers of the human organism [67].

The viral load required to infect a human host varies in the range of 10^2 to 10^3 viral particles, the diameter of each virion varies between 60 and 140 nm, and each aerosol produced during sneezing has diameters between 1 and 5 µm [68–70]. From a single cough reflex, an individual, with an average viral load of $7x10^6$ particles/mL, in the respiratory fluid, can transmit $4.4x10^3$ viral particles, and this number increases with increasing severity of the disease, that is, with increasing viral charge [71].

A study by van Doremalen *et al* [72] and Liu Y *et al* [73], report that the SARS-CoV-2 species has a viable stability of 72 to 196 hours (3 to 7 days) depending on the surface and origin of excretion (aerosol, urine or feces).

The average incubation time is around 4 to 7 days, reaching up to 14 days until the onset of symptoms [5,74]. After the incubation period, the patient may develop fever, non-purulent cough, laryngopharyngeal inflammation, dyspnea, hypoxemia, headache, myalgia, fatigue, olfactory dysfunction, diarrhea, lymphopenia, thrombocytopenia, acute myocardial ischemia, among others [22,64,74,75].

Unlike most viruses that mainly affect the upper respiratory tract, SARS-CoV-2 starts by lodging on and infecting the lower respiratory tract and pulmonary alveoli, giving rise to bilateral pneumonia, denoting opacity at the lower peripheral level of the lungs on images obtained by computed tomography (figure 2.2.1) [76].

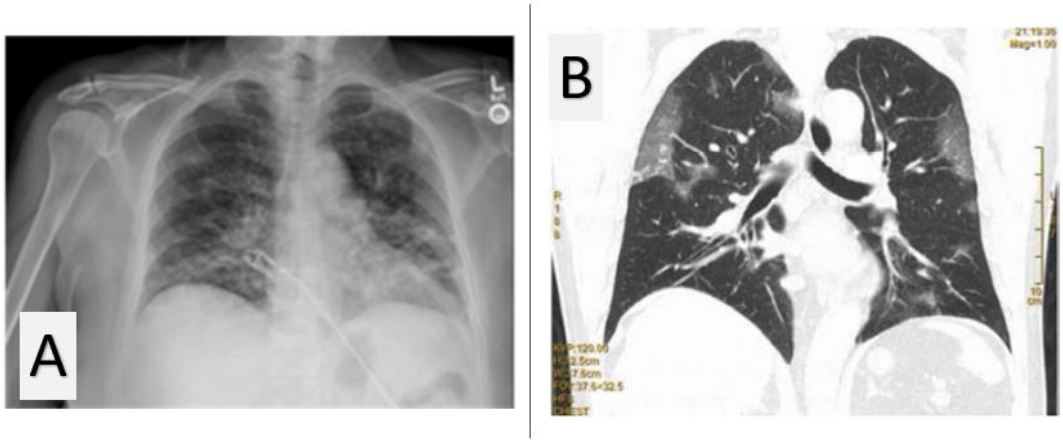


Figure 2.2.1 - COVID-19 pneumonia radiography and CT scans. (A) X-ray image showing a more prominent bilateral pulmonary infiltrate in the left lung. (B) Image obtained by computed tomography showing areas of increased opacity, islets of opacity are bilaterally distributed in the pulmonary zone. Adapted from [22] and [74].

Clinical manifestation of COVID-19 is apparent at all ages, with the prevalence being higher in the 30-79 age group (87%) as reported by Wu *et al* [77], an epidemiological study with 72 314 cases. Similar results are also observed, on a smaller scale, such as that by Dawei *et al* [78] where the median age is 56 (interquartile range of 42-68 years).

Individuals over 60 years of age, with a higher prevalence in males, and with various comorbidities tend to develop Acute Respiratory Distress Syndrome (ARDS), a complication of COVID-19, which requires hospitalization of the individual, in intensive care units, and is also one of the leading causes of death (figure 2.2.2) [79].

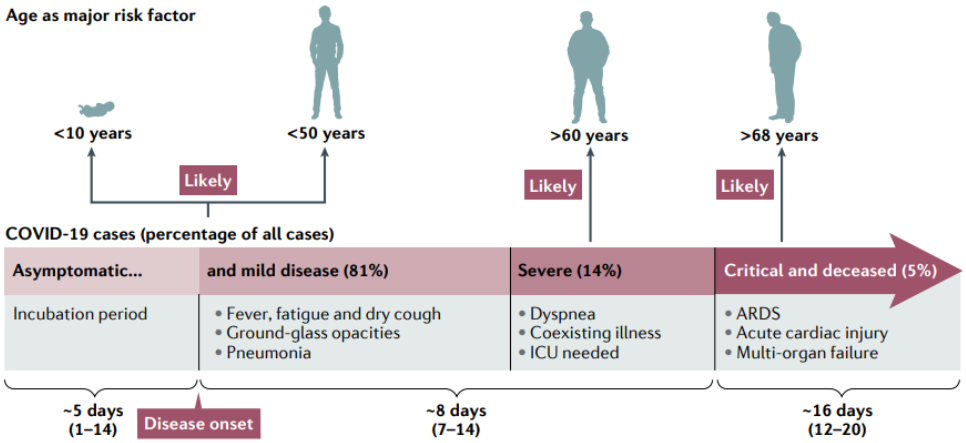


Figure 2.2.2 - Clinical features of COVID-19. COVID-19 clinical characteristics can easily be mistaken for a typical cold or pneumonia when the disease is set on a mild to severe case. Healthy patients tend to be asymptomatic, being the age a heavy risk factor. Age, and age associated co-morbidities are the major factors that lead to higher risk of critical development and in the worse cases, to death. Adapted from Hu *et al* [5].

Covid-19 events are directly linked to viral lung damage and cytokine release syndrome, an exacerbated inflammatory response that results from the production and secretion of pro-inflammatory cytokines and chemokines by immunocompetent cells in response to viral damage and cell sequestration caused by the virus (figure 2.2.3).

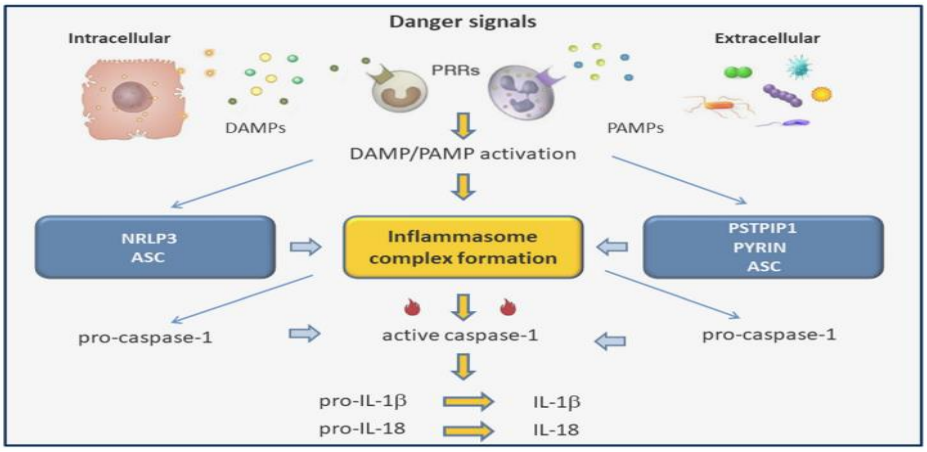


Figure 2.2.3 – Inflammasome complex formation. It's created during the pathogen-associated molecules (PAMP's) process, such in the case of SARS-CoV-2, and molecules associated with cell damage (DAMPs). PAMPs and DAMPs are recognized at the PRR (pattern recognition receptor) from APCs (Antigen-Presenting Cells). From the coupling between DAMP/PAMP-PRR an inflammatory cascade starts and activates the secretion of pro-inflammatory cytokines (like pro-IL-1β). In the case of the host cell, cytosolic receptors such as Nod-like receptors induce a specialized behavior of programmed cell death, apoptosis, releasing even more antigens capable of triggering a greater immune response. In normal conditions, these processes are regulated by the organism towards beneficial results, however, SARS-CoV-2 manages to dysregulate it. Adapted from Soy *et al* [80].

As the immune system identifies viral antigens through the normal process of recognition, antigen-presenting cells (APC's) process these antigens and present them to NK cells and cytotoxic CD8+ T cells at the MHC (major histocompatibility complex) level. This process activates both the innate and adaptive immune systems that cause the cytokine storm event. In some patients this process is dysregulated to the point of resulting in thrombotic events and multiple organ failure, and eventually causing death [80].

The renin-angiotensin system (RAS) is an important regulator of vascular pressure and coagulation, mainly composed by two axes, namely the ACE-Angiotensin II (Ang II)-angiotensin type 1 receptor (AT1R) axis, and ACE2-Ang-(1-7)-Mas axis. ACE regulates blood pressure and H_2O/Na balance by binding to AT1R and AT2R receptors (with greater impact and activity on AT1R). In addition to the above effects, it also has a role in

promoting inflammation, cell proliferation and vasoconstriction. ACE2 and ACE are structurally homologous, but have different biological functions, and the degradation of Ang II to Ang-(1-7) is mediated by ACE2. The combination of Ang-(1-7) and the MAS receptor antagonizes the ACE-Ang II-AT1R axis, which plays a role in lowering which plays a role in lowering blood pressure through vasodilation, decreased oxidative stress, and cell proliferation (figure 2.2.4) [79].

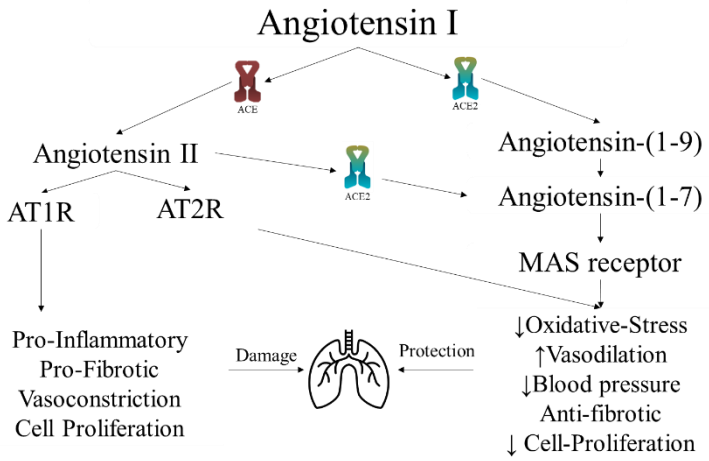


Figure 2.2.4 - Schematic diagram of the renin-angiotensin system (RAS). In the lung where the role of ACE2 receptor and ACE is visible as regulatory elements of RAS. The RAS pathway begins with the conversion of angiotensinogen to the angiotensin I (AngI) precursor through the action of the enzyme renin, produced in the kidneys. AngI is further recognized by the angiotensin-converting enzyme (ACE) cleaving it to angiotensin II (AngII). AngII is the peptide with the greatest influence on RAS. Its effect are mainly at the level of AT1R receptors. It also couples to the AT2R having the antagonistic effect. On the other hand, ACE2 receptors convert AngI and II into the inactive precursor Angiotensin (1-7) which, when interacting with the MAS receptor, has antagonistic effects to that of AngII [79,81].

As mentioned, the ACE2 receptor is expressed and widely distributed in the heart, kidney, adipose tissue, and muscle tissue, among others [35–39]. Being ACE2 the SARS-CoV-2 S protein receptor, there is a reduction in the availability of this receptor by viral sequestration, thus increasing the availability/conversion of angiotensin I to angiotensin II, and an over-expression of the signal resulting from AT1R. The function of the renin-angiotensin system is no longer balanced, which results in an exacerbated inflammatory response and the event of cytokine secretion that leads to lung damage, heart attack and pathologies located in tissues that express ACE2 receptors (figure 2.2.5) [79,81].

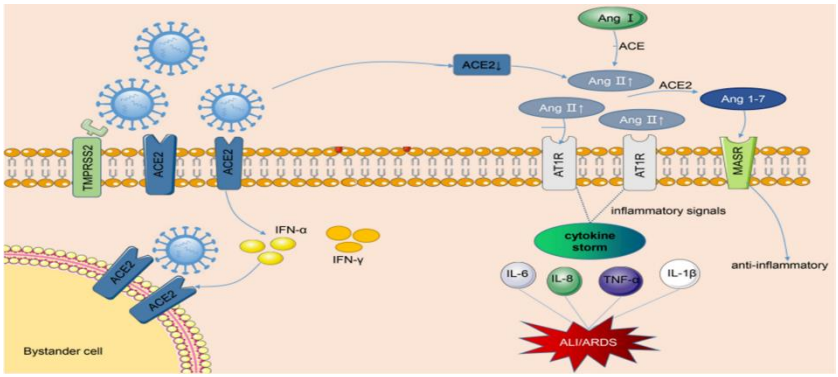


Figure 2.2.5 - Pathogenesis caused by SARS-CoV-2 viral infection in the renin-angiotensin system. The virion complexes with the ACE2 membrane receptor. After cell invasion, the natural antiviral pathway is activated, promoting the cellular secretion of interferons that have an apocrine effect with adjacent cells, promoting their expression of ACE2, creating new cellular targets for SARS-CoV-2. However, the levels of available ACE2 go down by sequestration, the levels of angiotensin II are promoted causing an uncontrolled inflammatory reaction with harmful effects at the level of the affected organ. Adapted from Gao *et al.* [81].

2.3. Metabolomics

In the field of OMICS (genomics, proteomics, metabolomics, metagenomics, and transcriptomics) the collective aim is to know how a pool of biological molecules affect a determined biological structure, function, and dynamics of an organism. The goals of this dissertation are within the field of metabolomics, which is the profiling of biological matrices focusing on the metabolites of the organism. Molecules studied on metabolomics, are small compounds (sugars, amino-acids, peptides, lipids, and others) produced by the endogenous metabolism of the organism of study. They are considered the chemical phenotypes derived from the genomics, transcriptomics, and proteomics, which provides a direct end-point, qualitative and quantitative, of biological processes under the influence of nutrition, environment, pharmaceutical alteration, and disease development [82–84]. Carrying the imprints of genomic, epigenomic, and environmental effects, the metabolome (defined as the complete metabolite collection found on a specific cell, organ or organism [85]) is considered the best fitting to study the causality of a variable of interest, for example, the effect of the infection by SARS-CoV-2 in the human organism [82,86].

Undeniably related to one's phenotype, metabolomics are already in use in many fields of medicine, such as, the determination of metabolite biomarkers for Parkinson's disease [87], type 2 diabetes [88], transplant rejection [89], shift in serum upon intense human activity [90], and many other as described by Minnie *et al* [91].

However, measuring the metabolome is in its core is a large and complex task. The human metabolome includes over 114,000 identified and predicted metabolites, including peptides and amino acids, lipids, nucleic acids, carbs, organic acids, biogenic amines, vitamins, minerals, xenobiotics, and drugs, as well as other chemicals that humans encounter [92].

It has been driven primarily by advanced analytic platforms, such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), having each platform its advantages and disadvantages (table 2.3.1).

Table 2.3.1 – Differences between the two major metabolic analytical platforms. Adapted from Attard *et al* [93].

	NMR	***MS***
Sample preparation	Minimal / non-destructive	Destructive
Reproducibility	High	Low
Quantification	Quantitative	Qualitative and quantitative
Relative sensitivity	Low	High
Metabolite coverage	Low	High

The initial workflow, generally, starts with an experimental design (selection a population of study, and the variable(s) to check), sample collection (solid samples as organ tissue, faeces, liquid samples as urine or blood cerebral-spinal fluid), preparation of the same sample passing through processes like homogenization of solid samples and extraction/precipitation of determined compounds for platform analysis suitability like the one mentioned (NMR and MS) which is the next phase. The final step, data analysis includes metabolite identification and data mining from the raw data, and data modelling using multivariate analysis and statistics to finally discover significant biomarkers. A validation step can be used to assure that the biomarker discovered has true potential, since it cannot be used in a clinical scenario until finishing necessary clinical validations about their specificity, sensitivity, and variation (figure 2.3.1).

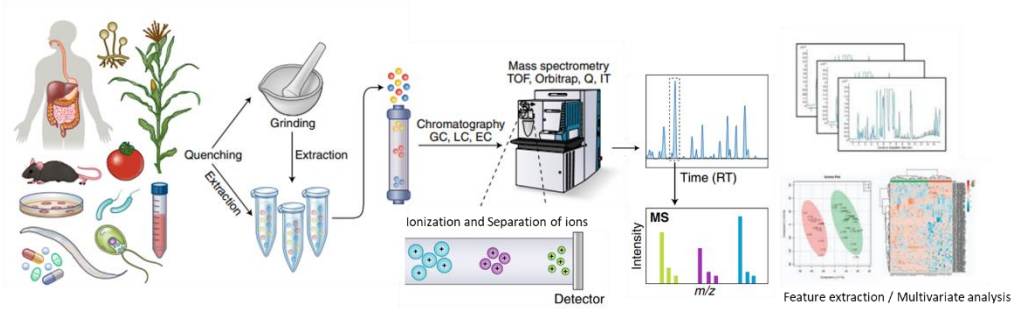


Figure 2.3.1 – Metabolomics work-flow. A typical metabolomic experiment involves several steps: cells, tissues and other biofluids can be isolated and extracted. Samples are then prepared for analysis. Commonly analytical platforms are coupled to other devises such chromatographic platforms, to separate the different components of the sample, reducing the sample complexity and rearrange the major sample molecules, facilitating the MS or NMR job to identify molecular profiles. The data retrieved can be then compared to large databases. After that, statistical and multivariate algorithms are used to verify the significance of those molecules, and then the molecule(s) that pass through the sorting are tested as a differential metabolite(s), which is, the ability to distinguish normal individuals from patients with the variable of study for example. Adapted from Alseekh *et al* [94] and Li *et al* [87].

The study of metabolomics allows comprehensive quantitative/qualitative analysis of all metabolites in a biological sample. It is divided into two types of approach, the targeted approach, in which the assessment of a certain set of pre-established compounds is conducted using certain standard compounds, and the untargeted approach, which broadly investigates all metabolic profiles present in each sample using high resolution methods (table 2.3.2). While an untargeted analysis will detect any molecule that ionizes within a certain range of mass values, a targeted analysis provides better quantification, typically using internal standards and specific conditions of the mass spectrometer coupled to a chromatography equipment (set of equipment widely used in metabolomics) sacrificing in this case the number of molecular characteristics existing in each sample compared to untargeted analysis [82,95,96].

Table 2.3.2 – Main differences between Untargeted and Targeted Metabolomic Analyses. Adapted from Schrimpe *et al* [97].

Untargeted analysis	Targeted analysis
. Reproducibly measure as many metabolites as feasible . Hypothesis-generating . Semi-quantitative (chromatographic peaks reported are not concentrations) . Feature identity is not priori known . Correlated to databases/libraries	. Focus on a small number of molecules . Hypothesis-driven / proofer . Absolute quantification . Identification already known . Correlated to reference standards

The initial approach, when the objective is to discover potential new biomarkers, is the *untargeted* analysis, a extensive scanning for all molecules present in the sample. The patterns identified can be then used to generate potential hypothesis for a biomarker, focused more on a niche of metabolites. Ultimately that biomarker can be tested in a targeted way [97].

COVID-19 diagnosis is usually based on the qualitative detection of viral RNA, with most studies focusing on the clinical and epidemiological characterization of patients. As Huang *et al* reported, about 85% of patients infected with COVID-19 show improvement and are discharged from the hospital, and about 68% of patients do not need to be admitted to an intensive care unit, as corroborated by data from Wang *et al*. and Guan *et al* [64,75,77,78,98,99].

However, about 15-20% of patients suffer from more severe symptoms such as severe respiratory disease and require oxygen therapy or other interventions, including mechanical ventilation. These patients are classified as clinically severe, and are mostly classified according to clinical characteristics such as respiratory rate, partial oxygen saturation, and/or mean oxygen saturation [77]. However, patients who manifest these conditions are

already in a severe clinical state, so the therapeutic intervention must be urgent and effective, otherwise the patient can die quickly.

Understanding the biochemical profile and its changes in a healthy or even asymptomatic individual regarding the manifestation of COVID-19 is a challenging topic. The same problem can be raised regarding the probability of survival and severity of individuals who are infected with SARS-CoV-2 and who are or are not undergoing therapy. Being able to respond to the problems put above is to achieve a greater speed of response and prevent negative outcomes such as death and the development of aggravated comorbidities caused by SARS-CoV-2. Through a metabolomics analysis, it is possible to characterize the molecular profile of an individual at any time and to understand and predict the impact of the disease on the body. Metabolomic studies allow not only the study infection's effects by the infectious agent, but also to measure the response of the host against the pathogen, and therefore, these studies can provide a set of markers, complementary, and of significant use for rapid tests, parallel to antibody and PCR testing, for confirmation of SARS-CoV-2 infection, severity, and potential COVID-19 disease outcome.

Given the importance of this area, several studies have applied metabolomics to investigate COVID-19 disease in human patients, which can be seen in table 2.3.3.

Table 2.3.3 - Overview of results and models from published studies using metabolomics to characterize COVID-19 patients.

Reference	Patient group	Sample preparation	Methodology / Instrumentation	Result	Metabolite/Pathway evidenced
Roberts *et al* [100]	Serum samples from 120 COVID-19 patients	Protein precipitation 3:1 (V: V) of methanol with sample (300 µL:100 µL), vortexed and centrifuged at 18000g x15 min. at 4ºc. From the resulting supernatant, it was evaporated and resuspended in 40 µL H_2O.	Untargeted metabolomics using UHPLC-MS/MS. Data obtained from LC-MS exported to Compound Discoverer 3.1, where the compounds obtained were filtered and selected from 20 compounds that were used to make the predictive models	Identification of prognostic biomarkers of COVID-19 severity and outcome	• Pyrimidine • Tryptophan • Aspartate • Asparagine
Doğan *et al* [101]	Serum samples from 41 COVID-19 patients and 44 controls	Samples were homogenized and mixed with acetonitrile (500 µL: 1000 µL, respectively) placed at 4°C, 10min. and centrifuged at 16900g x 5min. 500 µL of the supernatant was transferred to new microtubes. An internal standard collocated to improve data quality.	Untargeted metabolomics using LC-QTOF-MS	Identification of endogenous metabolites in COVID-19 patients and healthy controls	• Glutamine • Glutathione • Leukotriene D4 (LTD4) • Purines
Shen *et al* [102]	Serum samples from 65 COVID-19 patients, 25 non-COVID-19 patients and 28 healthy controls.	Whole blood samples centrifuged at 1500 g x 10 min. the serum obtained was collected into new centrifuge tubes and stored at -80°C. For metabolomic analysis, ethanol was added to the samples to deactivate potential viruses, and subsequently dehydrated. To the deactivated serum (100 µL) was added 300 µL of methanol, to which the proteins were removed by centrifugation.	Untargeted metabolomics using UPLC-MS/MS (RP and HILIC in ± mode) and Proteomics. From the molecular data obtained, the authors developed a machine learning model to identify severely ill patients according to a range of selected molecules.	Detection of metabolic and proteomic changes in the serum of severe COVID-19 patients, which can be used in the selection of potential biomarkers for assessment of disease severity.	373 altered metabolites in COVID-19 patients, 203 involved in severity, 80 related to: • Complement system • Macrophage function • Platelet degranulation • Lipid metabolism
Su *et al* [103]	Collection of peripheral blood plasma and immune mononuclear cells samples from 139 COVID-19 patients, and 258 healthy subjects.	Study samples collected in EDTA tubes. Plasma was collected by centrifugation at 800g x10min., 4°C, aliquoted and stored at -80°C. mononuclear cells were extracted using the Ficoll gradient method. Serum processing effectuated at Metabolon (Morrisville, NC, USA) facility, with no details	Untargeted metabolomics using UHLC-MS/MS	significant shift between severe and moderate patients, in which the prominent level of the inflammatory signal is accompanied by the loss of specific classes of metabolites and metabolic processes.	776 metabolites included for analysis, involved in the metabolic pathways of: • Amino acids •Nucleotides •Carbohydrates •Lipids •xenobiotics

Thomas *et al* [104]	Serum samples from 33 COVID-19 patients and 16 controls.	Serum obtained by passive sedimentation of blood stored at 4ºC. 20 µL was diluted with 130 µL of H_2O, to which 600 µL Chloroform/Methanol (2:1, V/V) was added, and then vortexed. The samples were incubated for 5 minutes at 4ºC, vortexed again (5 seconds), and centrifuged at 14000g x10min., 4ºC. the supernatant resulting from this process was isolated and used for metabolomic analysis.	Targeted and untargeted metabolomics using UHPLC-MS. Acquired data were analysed with Maven and Compound Discoverer 2.1 programs.	Observational characterization of sera from COVID-19 patients.	• Amino acids (focus on tryptophan) • Lipids • Urea cycle • Glucose Cycle
Páez-Franco *et al* [105]	Serum samples from 46 severe and 19 mild COVID-19 patients, and 27 PCR negative controls for SARS-CoV-2.	Serum sample was spiked with an internal standard. It was added chloroform-methanol (1:3). Incubation of the mixture with sample for 20 min. 20 ºC, and centrifuged 14,000 rpm for 15 min. The resulting supernatant was dried and resuspended in methoxyamine and incubated for 90 min at 37 °C in a shaking incubator. MBSTFA+1% TMCS was added to each sample and incubated for 30 min at 37 °C. On microliter of the final mixture was submitted to GC/MS system.	Untargeted metabolomics using GC-MS. Data obtained from mass spectrometry equipment converted to. mzdara type using the Agilent ChemStation software. Feature detection, spectral deconvolution and peak alignments done using the Mzmine2 software. Identification of compounds by comparison with the NIST (National Institute of Standards and Technology) spectral library.	Identification of the main serological changes associated with severe and mild illness of COVID-19 patients and synthesis of a metabolic profile for these two features of the same disease.	• Krebs cycle (Warburg effect) • Catabolism of amino acids (threonine, glutamine, and glutamate)

2.4. Non-Conventional Metabolomics

As described above, metabolomic analysis can be time consuming, expensive, and complex, as it involves many stages and multiple steps for sample suitability for a given analysis platform, and most of the time, the sample is lost for a destructive analysis technique such as mass spectrometry.

New methods to quickly and easily assess the effect of one variable on another and, extrapolating to hospital context, the effect of the infection by SARS-CoV-2 in in one individual, are highly useful to predict the outcome and severity caused by COVID-19.

Platforms such as FTIR (Fourier Transformed Infrared) analysis are an emerging technique that allow sample analysis (biological tissue, biofluid and even gases). FTIR analysis, uses radiation in the field of the infra-red (range between 780 nm to 1 mm).

FTIR is placed in the category of vibrational spectroscopies, as the energy produced in the IR field is not powerful enough to cause physicochemical changes in the matter it passes through, such as ionization and electronic transition of electrons. Instead, infra-red radiation causes distortions on the molecular bonds, changing their angle, length and other phenomena related to atom bonds, as shown on figure 2.4.1.

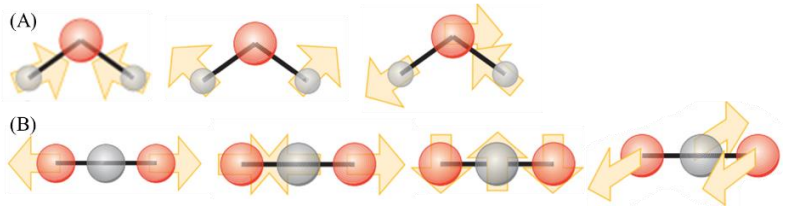


Figure 2.4.1 – Every molecule has several modes which they can vibrate at. (A) A nonlinear molecule such as H_2O has a 3N-6 different ways that it can vibrate (being N the number of atom nuclei present on the molecule). (B) linear molecules such as CO_2 has 3N-5 (being N the number of nuclei atoms on the molecule). The vibrational modes pass through stretching, symmetric or antisymmetric, depending on if the O-H bond are in or out of phase, respectively. It can be as well a bending vibration, if the angle between the H-O-H is changing, also called the scissoring motion. In the case of the molecule of CO_2, it has 4 normal modes of vibrations (name given to each way which a molecule can vibrate), symmetrical stretch where the O-C-O bond elongates in opposite directions, maintaining the 180º angle, antisymmetrical stretch where one O-C bond is shorter than the other, still maintain the 180º angle. And finally, the Bend, where the O-C-O 180º angle is changed. Adapted from [106]

Infrared spectroscopy corresponds to the observation of the vibrations described, and with those vibrational observations, one can enquire and search for evidence of a functional molecular group, and ultimately uncover the compound of study. FTIR uses this approach, as demonstrated on figure 2.4.2.

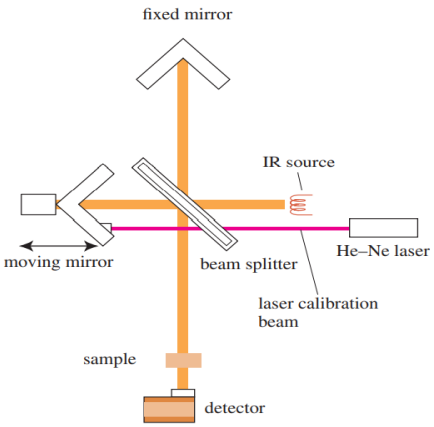


Figure 2.4.2 – Diagram of an inside FTIR spectrometer. The IR light is directed from the incandescent source to a beam splitter, positioned at a 45° angle. A portion of the beam is reflected at a straight angle after passing through the beam splitter. The reflected beam hits a fixed mirror, whereas the transmitted beam hits a moving mirror at a constant speed. The beams recombine at the beam splitter after returning from the mirrors. The moving mirror's beam has travelled a different distance than the fixed mirror's beam, and the two beams combine to form an interference pattern known as an interferogram. This interferogram, which contains all frequencies at the same time, travels through the sample compartment to reach the detector. A Fourier transform, a typical computer procedure, changes the time domain spectrum to the frequency domain spectrum, allowing us to measure the degree of absorption as a function of wavelength (or frequency). Adapt from Wade *et al* [107].

Each bond group (bond formed between two atoms) has a vibrational transition with a well delimited and distinct energy, and molecules absorb only at certain wavelengths and frequencies. It is unlikely that IR spectra of two compounds will be the same (except for enantiomers), making each spectrum of each molecule a unique representation on the IR field, and thus it can be considered that a molecule IR spectrum is its fingerprint, and this is what will allow one to determine a composition of a bio sample.

The major advantage of this technique is its sensitivity, equipment with low energy consumption and low time of "sample reading" (in the order of the seconds of each scan). Multiple scans for the same sample can be carried on improving data quality, since this technique does not destroy the sample (table 2.4.1).

Table 2.4.1 – Main features of IR spectroscopy. Adapted from Wade *et al* [107].

Strengths	Limitations
. Identification of functional groups. . It can provide conclusive proof that two unknown compounds are the same or different. . Highly reproducible and low-resource usage . Minimal sample preparation needed	. Unable to offer insight into a molecule's skeletal structure. . In some cases, it's not possibly to certain of the functional group due to minimal differences in the spectra . In complex mixtures such as biofluids, where thousands of different molecules are present, is complex to extract the exact number of molecules present or what molecules are present without additional information provided by other techniques.

In table 2.4.2 it can be observed the main vibrational frequencies in FTIR spectroscopy.

Table 2.4.2 - Vibrational frequencies in FTIR spectroscopy for serum. Adapted from Wood *et al* [108], Movasaghi *et al* [109], Cakmak *et al* [110].

Wavenumber (cm^{-1})	Band Assignment
3600–3500	Amide A, proteins, N-H bond stretch
3350	N-H bond stretch
2956	Acyl chain, Lipids, CH_2 and CH_3 vibration
2922	
2874	
2852	
1740	Lipids , C=O vibration
1650 - 1655	Amide I, C=O stretching, peak of nucleic acids due to the base carbonyl stretching and ring breathing mode
1635	Amide I
1545–1530	Amide II
1300–1250	Proteins, amide III
1185	A-DNA
1060	DNA, RNA
996	RNA

Chapter 3 – Materials and Methods

3.1. Biological data acquisition

To analyse the impact and outcome from Coronavirus disease, biological samples were collected from Centro Hospitalar Universitário Lisboa Central – Hospital São José, intensive Care Unit (ICU). Blood sample collection started in November 2020, within PREMO Project - *Predictive Models of COVID-19 Outcomes for Higher Risk Patients towards a Precision Medicine* -financed by FCT. 3865 samples were collected until August 2021, from 521 patients.

Samples were collected by the health professionals present at the medical unit at 8 a.m. to tubes without anticoagulant (Figure 3.1.1), transported to /and processed at the Instituto Superior de Engenharia de Lisboa at the Engineering and Health Laboratory on the same day and as soon as possible. All biological content in the transport process was kept at 4 ºC until processing.

Peripheral blood samples were centrifuged at 3000 rpm, 10 min., room temp. (Mixtasel, J.P. Selecta ®) to separate serum from the erythrocyte layer.

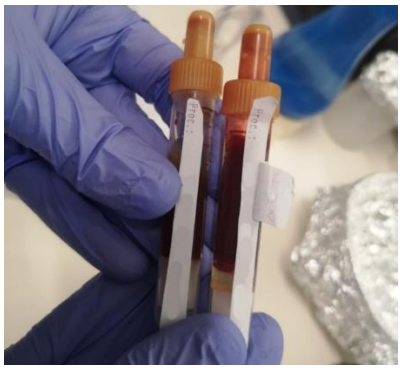


Figure 3.1.1 - Properly identified peripheral blood collection tubes (censored for ethical reasons).

After centrifugation, the supernatant corresponding to the serum fraction was extracted and aliquoted in 1.5 microtubes. Each aliquot was stored at temperatures of -20ºC and -80ºC respectively until further analysis (Figure 3.1.2 - Stored serum aliquotsFigure 3.1.).

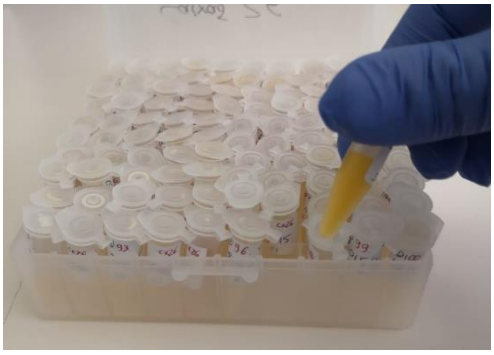


Figure 3.1.2 - Stored serum aliquots.

Blood serum contains many thousands of molecules that can provide biologically relevant information. The final goal of this work is to use metabolomics to find a set of biomarkers that can differentiate different health states between patient groups, namely, the outcome and patient condition upon infection by SARS-CoV-2. To accomplish this objective, two analytical techniques were use, FTIR (Fourier-transform infrared spectroscopy) and UHPLC-MS (ultra-high performance liquid chromatography coupled to mass spectrometry). Figure 3.1.3 shows the workflow of the experiment ran.

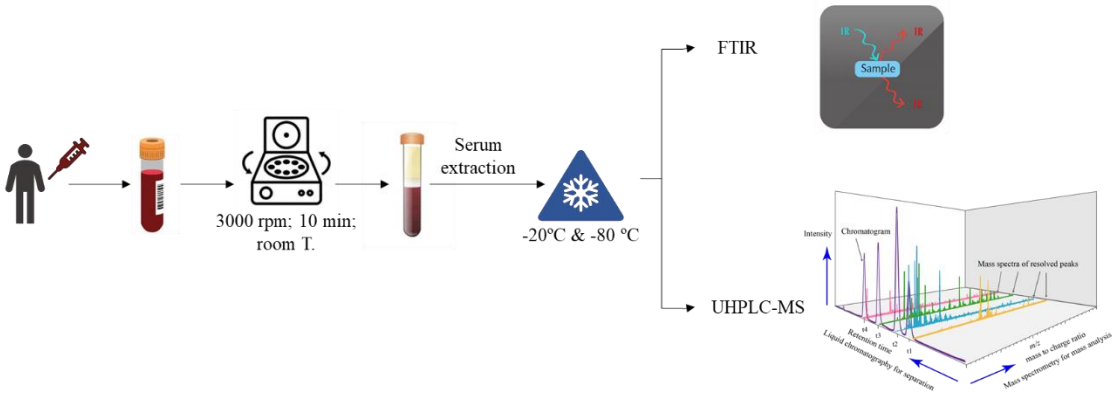


Figure 3.1.3 - Schematic view of the procedure executed. The blood from the patients is collected directly from the ICU stored in collection-blood tubes without anticoagulant to facilitate the blood clotting process in which is easier to extract the serum through centrifugation. The serum is aliquoted in at least 2 microtubes with a volume of 2 mL and stored each volume at -20 ºC and -80 ºC refrigerator for further analysis (FTIR and/or LC-MS).

For FTIR and MS analysis the chosen serum samples (criteria of patient inclusion discussed later) were processed ad follows:

1. Preparation of distinct solvents for extraction. Two approaches were used, one with methanol (Fisher, OPTIMA® LC/MS grade) and another with a mixture of acetonitrile (Fisher, OPTIMA® LC/MS grade):methanol (Fisher, OPTIMA® LC/MS grade):water (Fisher, OPTIMA® LC/MS grade) (at 2:2:1 V/V) was prepared and kept at -80 ºC.
2. Solvents were aliquoted into 2 mL microtubes for protein extraction.
3. Chosen samples were thawed and kept on ice throughout handling.
4. From each sample, 75 µL were pipetted to the microtube with the solvent, at a ratio of 3.5:1 V/V solvent to sample.
5. The microtubes with the sample plus the solvent were centrifuged at 18 000 x g, 15min at 4 ºC (MIKRO 22R Hettich Zentrifugen).
6. The resultant supernatant was carefully isolated from the formed pellet and transferred to a new microtube.
7. 75 µL of newly collected supernatant was stored for FTIR analysis and the leftover volume was stored for UHPLC-MS analysis.

Protein precipitation (PPT) is one of the most popular techniques used to prepare biological fluid (bio-fluid) samples for LC/MS/MS research analysis. By applying protein precipitation to bio-fluid samples, proteins can efficiently be removed from the serum's matrix.

3.2. FTIR Spectroscopic analysis

For FTIR analysis the samples were placed in a zinc selenide microplate (FTIR able) with 96 or 384 wells (figure 3.2.1) depending on the number of sample replicas and extraction that was dealt (explained below).

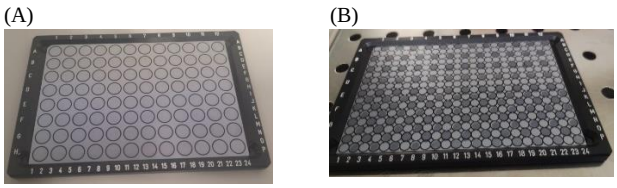


Figure 3.2.1 - ZnSe microplates used for this technique. (A) 96 well ZnSe microplate, with each well holding a volume capacity of 25 µL; (B) 384 well ZnSe microplate with a volume capacity of 5 µL each well.

The microplates with the samples were then submitted to evaporation by desiccator/vacuum chamber, for 2h 30min, to remove all solvent as show in figure 3.2.2.

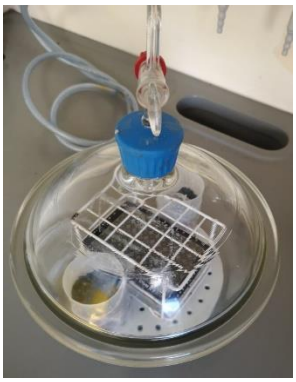


Figure 3.2.2 – Microplate with loaded samples in the wells placed on a rack in a desiccator. In the image is possible to observe two cylinders containers with silica crystals that helps the evaporation process.

The next step was to analyse the dehydrated samples in a Fourier-transform infrared spectrometer (Vertex 70, HTS-XT, Bruker). For each sample loaded in the microplate well a spectrum was acquired. For each spectrum 64 scans, with a resolution of 2 cm^{-1}, in transmission mode, were collected. The spectral acquisition comprehended wavelengths between 400 and 4000 cm^{-1}. The first well of the microplate did not contain sample and spectra was acquired and used as background, according to the HTS-XT manufacturer instructions.

All data collected was analysed with the OPUS software, where spectra were pre-processed employing atmospheric correction (that compensates atmospheric H_2O and CO_2 at acquisition) and Baseline Correction (rubber band baseline type, positive peak direction, with background subtraction, used often in IR spectra to adjust the spectral offset). Pre-processed data was then uploaded to Excel for matrix organization and then loaded on The Unscrambler® X for matrix transformations, such as normalization (is a mean to scale the matrices data to retrieve all spectral data to the same scale), and derivation like Savitzky-Golay (also known as SG derivate, used to perform data smoothing through a process called convolution). Workflow from spectra acquisition to data pre-processing can be observed in Figure 3.2..

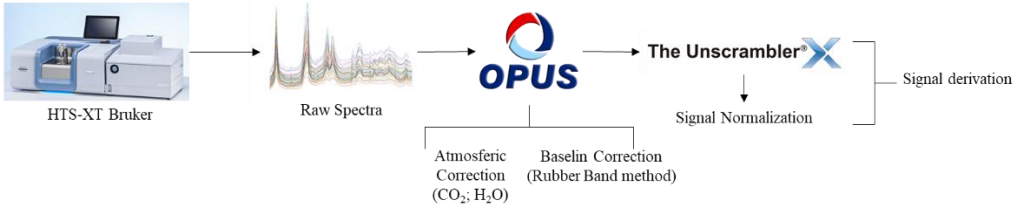


Figure 3.2.3 - Workflow process done on FTIR analysis. Dried biological samples are read at the FTIR spectrometer. Raw data is loaded into OPUS software for primary pre-processing. Signal normalization and/or derivation is carried out on The Unscrambler® X where the data is smoothed and treated into more amenable signals.

In this study, spectra processing is followed by a multivariate plot analysis, Principal Component Analysis (PCA), used to reveal the hidden structure within large data sets. It provides a visual representation of the relationships between samples and variables and provides insights into how measured variables cause some samples to be alike, or how they differ from each other. It is a bilinear model, that projects all variables from a dataset unto a smaller number of newly created variables, named principal components (PC) by default equal or less than seven. Large data tables can contain a lot of information to consider from a human perspective, or data can be hidden in a complex and not easily interpreted way. Examples of these tables are spectroscopic data, chromatographic data, and the large data sets generated by systems biology research platforms, in particular, metabolomics. A few examples of the desired signal pre-processing can be observed in figure 3.2.4.

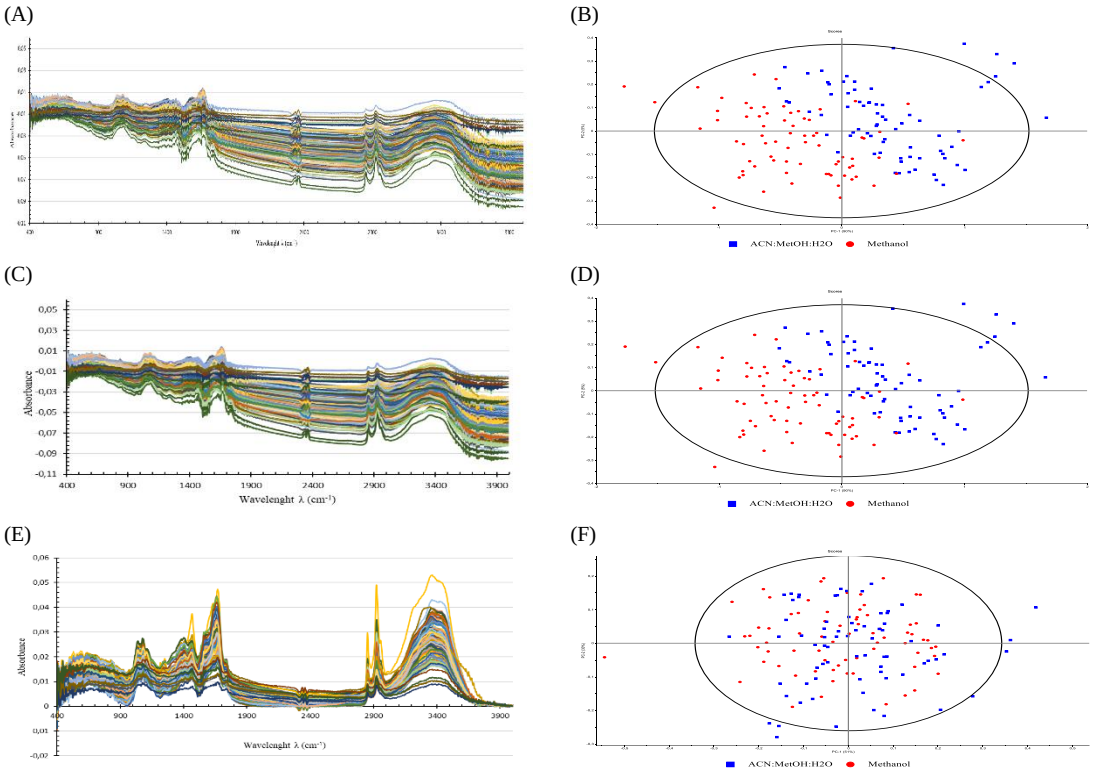


Figure 3.2.4 - Spectra obtained from FTIR spectrometer from the solvent extraction assay (explained in chapter 3.4). Spectra in the left present absorption intensity vs. wavenumber, as recorded on the HTS-XT FTIR. (A) Raw spectra with no pre-processing. (C) Spectra submitted to Atmospheric correction executed in OPUS software. (E) Spectra submitted to Atmospheric Correction followed by Baseline, Rubber band method, correction executed on OPUS software. B, D and F are the PCA's of spectra A, C and D respectively, and presented to facilitate a visual insight of the spectra shown. Each dot in the PCA represents a sample, extracted using methanol (red dots) and ACN:MeOH:H_2O (blue dots).

Normalization of data, such as FTIR signal data, has a purpose to "scale" data into an order to get all signal on the same scale. In the present work the more often normalization methods used are the Unit Vector Normalization (UVN) and the Maximum Normalization methos. As the name implies, unit vector normalization transforms the matrix sample data into unit vectors with a norm of one, keeping all the properties in check, which is useful in some pre-processing in some pattern recognition techniques. The other one, maximum normalization, divides each row of data by its maximum absolute value, retrieving a spectrum that has a maximum value of 1 to the maximum value of each row, the remaining values become a percentage of that given maximum value. Examples of this normalization techniques can be visualized in Figure 3.2.5.

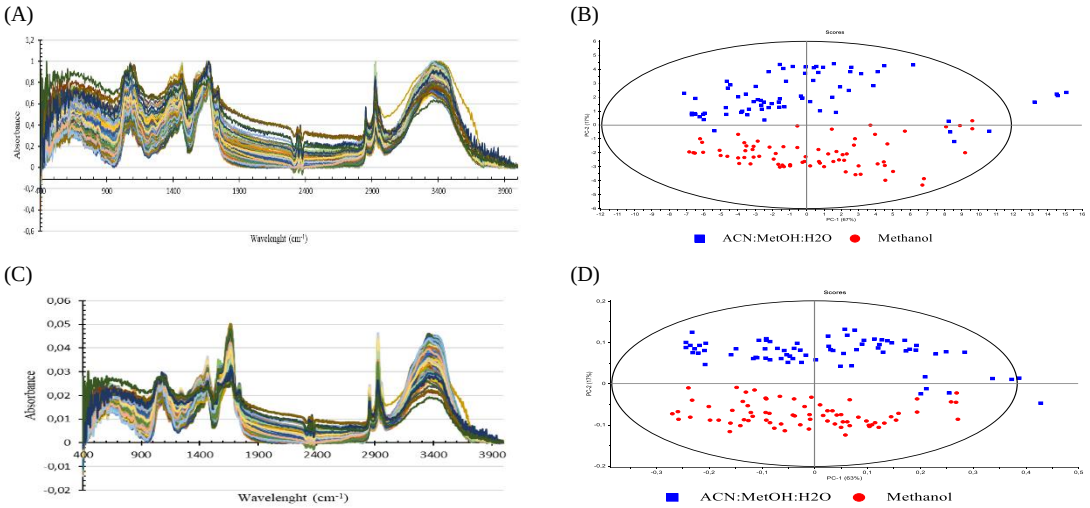


Figure 3.2.5 - Spectra from the different solvent extraction assay. (A) Spectra resultant from the application of Atmospheric correction to eliminate potential CO_2 and H_2O noise, Baseline Correction, Rubber band method, and a posterior Maximum Normalization executed on The Unscrambler® X software. (C) Spectra resultant from the application of Atmospheric correction, Baseline Correction and a posterior Unit Vector Normalization. B and D are the corresponding PCA's for A and C, and here disposed for a better observation of the impact of each normalization done in the data.

As mentioned earlier, the use of signal derivation is one method to smooth data, as the derivative is simply the measure of the slope of the spectral curve, and so, it is not affected by noisy elements of the spectra as baseline offsets. For this reason, derivatives are an effective way to remove this element from spectral data. However, one drawback of this method is that peaks in raw spectra usually become zero-crossing points, what could be difficult to interpret. To bridge this undesired event, Savitzky-Golay (SG) method is used in this work. It consists in using a localized segment of the spectrum and apply the derivative for that specific data, avoiding noise enhancement and allowing data smoothing. Figure 3.2.6 shows an SG smoothing method application.

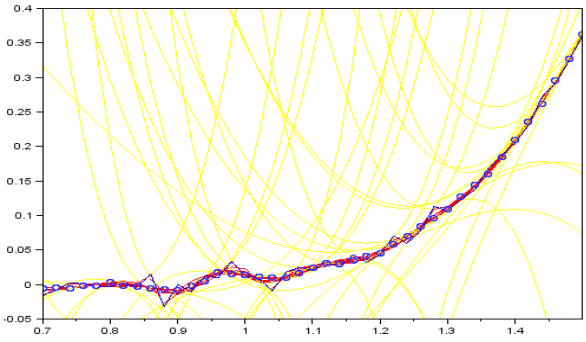


Figure 3.2.6 - Savitzky-Golay application. As observed multiple derivatives (yellow lines) are being applied tangent to raw data (in blue) and the segment of that derivate (red line) is being used to fit the new data sub-set that will be the new data (blue circles).

Examples of the applications of the methods explained above are show in Figure 3.2.7.

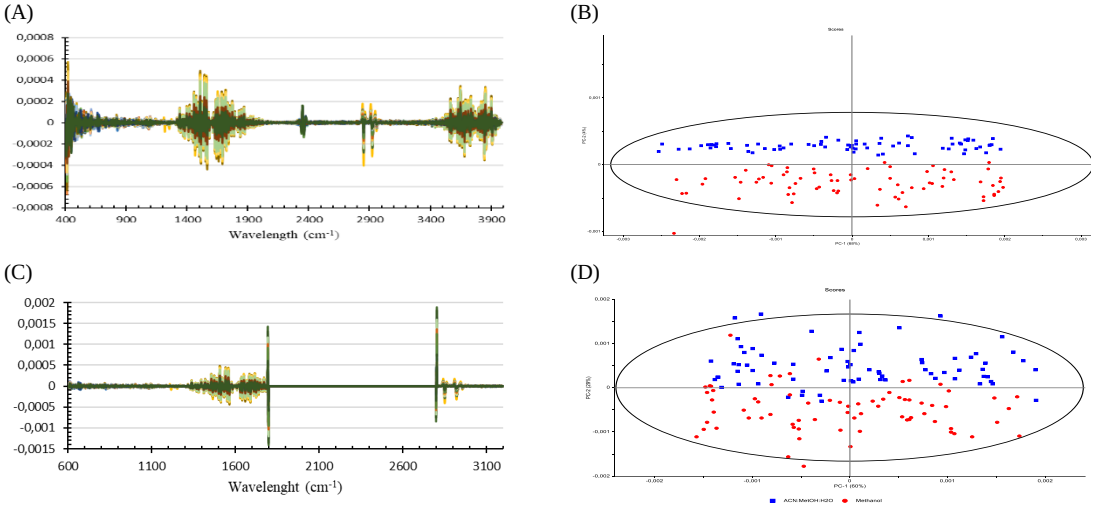


Figure 3.2.7 - Spectra from the different solvent extraction assay. (A)Spectra data relative to extraction assay, which suffered Atmospheric correction on OPUS software followed by Savitzky-Golay Transformation with a derivative Order of 2, Polynomial Order of 2, with 15 smoothing points. (C) Is the same spectra present in (A) but with a wavelength cut applied, 800-1800-2800-3600 cm^{-1}.

3.3. UHPLC-MS Analysis

Alongside with FTIR analysis, samples were analysed by ultra-high performance liquid chromatography coupled to mass spectrometry. High-resolution mass spectrometry (HMRS) analyses were performed on an Elute UHPLC system (Bruker Daltonics GmbH) coupled with a QqTOF Impact II mass spectrometer with an electrospray ion source (Bruker Daltonics GmbH & Co.) (UPLC-ESI-HRMS/MS). Data acquisition and calibration were performed using the software Data Analysis 4.1, 4.4 and 4.5. Workflow of mass spectrometry analysis can be observed in the following figure 3.3.1.

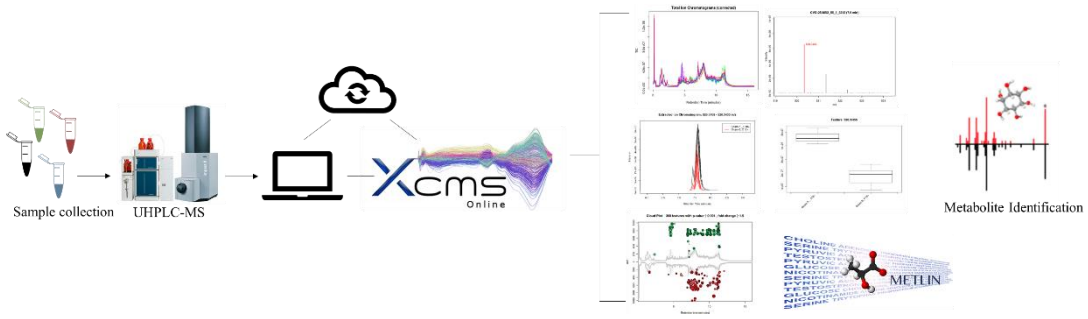


Figure 3.3.1 - Normal workflow followed during MS analysis. Data from UPLC-ESI-HRMS/MS are converted to a compatible format and uploaded into XCMS online servers, where group comparison are made. XCMS treats data in an automated way, such as, chromatographic peak detection and retention time alignment, as well mass spectra matching [111]. From there various plots are generated, like, boxplot for the *m/z* intensities of each group and cloud plot for significant features detected. XCMS uses METLIN open-source database to identify features into predictive metabolites [112–115].

Metabolites were separated by reverse-phase chromatography on a Luna 2.5 µm C18(2)-HST column (100Å, 150 x 2 mm, Phenomenex, Cat# 00D-4446-B0) at a constant temperature of 40 °C, using a gradient elution at a flow rate of 250 µL/min (mobile phase A: 0.1% formic acid in water; mobile phase B: 0.1% formic acid in acetonitrile): 0.0-0.5 min 0% B, 0.5-1.5 min 0 to 20% B, 1.5-4.0 min 20 to 60% B, 4.0-6.0 min 60 to 100% B, 6.0-9.0 min 100% B, 9.0-10.0 min 100 to 0% B, 10.0-15.0 min 0% B. For hydrophilic interaction chromatography (HILIC), a XBridge BEH Amide XP Column (130Å, 2.5 µm, 150 x 2.1 mm, Waters, Cat# 186006724) was used at a constant temperature of 40 °C. With a flow rate of 250 µL/min, a gradient elution of 10mM ammonium acetate in water containing 0.1% acetic acid (A) and 10mM ammonium acetate in acetonitrile containing 2% of water and 0.1% acetic acid (B) was applied: 0-2 min 90% B, 2-6 min 90 to 70% B, 6-9 min 70 to 30% B, 9-13 min 30% B, 13-18 min 30 to 90% B, 18-22 min 90% B.

MS acquisition parameters were set as follows: capillary voltage 3 kV (ESI+) or 4.5 kV (ESI-), end plate offset 500 V, nebulizer 4.0 bar, dry gas flow 8.0 L/min, dry heater temperature 200 ºC. Spectra acquisition was performed with an absolute threshold of 25 counts per 1000. For auto MS/MS data acquisition, capillary voltage was set at 4.5 kV (ESI+) or 3.5 kV (ESI-), with an end plate offset of 500 V, nebulizer pressure of 4.0 Bar, dry gas flow of 8.0 L/min and a heater temperature of 200 °C. Spectra acquisition was performed with a threshold of 20 counts per 1000, a cycle time of 3.0 seconds with exclusion after 3 spectra and release after 1.00 min. Internal calibration was achieved with a sodium formate/acetate solution introduced to the ion source *via* a 20 µL loop at the beginning of each analysis. Calibration was then performed using high-precision calibration mode (HPC). All acquisitions were performed with a *m/z* range from 50 to 1300 a.m.u. with a 3 Hz spectra rate. 3 full scans and 1 auto MS/MS scans were performed for each sample using both positive and negative ionisation mode.

Acquired MS data were processed by Data Analysis (version 4.1, 4.4 and 4.5) (Bruker Daltoniks). Raw data were converted to mzXML using ProteoWizard MSConvert [116] (figure 3.3.2).

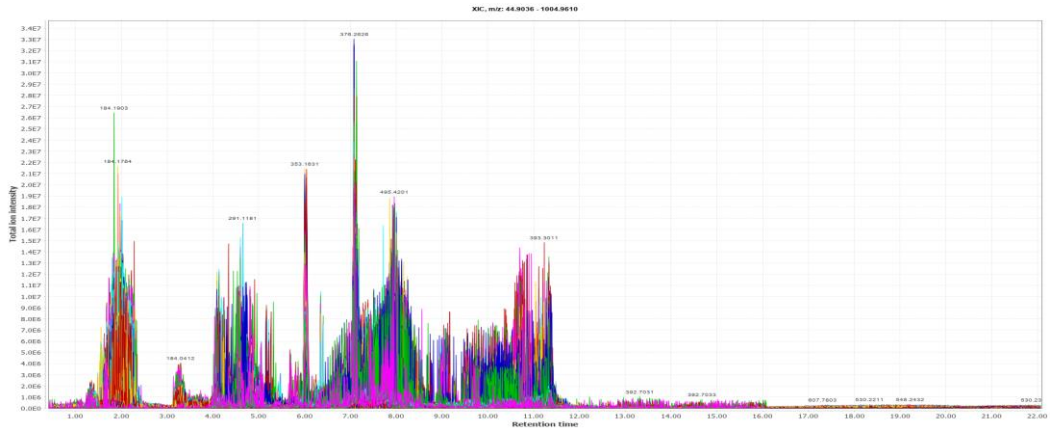


Figure 3.3.2 - Example of an Extracted Ion Chromatograms (XIC). Like Total Ion Chromatograms (TIC) where all intensities of each spectre scan are summed up and reflected on the graphic as a total intensity at a given retention time, XIC graphic give us, for a given retention time, the most intense observation of that spectral scan and returns it as total intensity. In the Y axis is the total ion intensity and X axis the retention time (in minutes). This XIC is the representation of the signal acquired from the extraction method assay [117]. Platform used for visualization: MzMine 2.53 [115].

Raw UHPLC-MS data were converted to a more amenable format, mzXML, and uploaded to XCMS online servers [113], where all data analysis (namely sample/group pairwise comparison). Multimodal analysis independent of acquisition model, and global metanalysis, were performed. Parameters used for pairwise comparison between subgroups is shown in table 3.3.1.

Table 3.3.1 - Parameters used in XCMS Online analysis for patients' samples untargeted metabolomics data process.

Step	Option	Value
Feature detection	method	centWave
	ppm	10
	minimum peak width	5
	maximum peak width	20
	mzidff	0.01
	signal/noise threshold	0.01
	integration method	1
	prefilter peaks	3
	prefilter intensity	100
	noise filter	100
Retention time correction	method	centWave
	profstep	1
Alignment	bw	5
	minfrac	0.5
	mzwid	0.015
	minsamp	1
	max	100
Statistics	statistical test	Welch *t*-test for unequal variances
	perform post-hoc analysis	True
	p-value threshold (highly significant features)	0.001
	fold change threshold (highly significant features)	1.5
	p-value threshold (significant features)	0.05
	value	into
Annotation	ppm	5
	m/z absolute error	0.015
	search for	Isotopes + adducts
Identification	adducts	$[M+H]^+$ (ESI+); $[M-H]^-$ (ESI-)
	ppm	5
	BioSource	Human
	pathway ppm deviation	5
	significant list *p*-value cut-off	Auto

XCMS-generated data (figure 3.3.3) was returned and analysed.

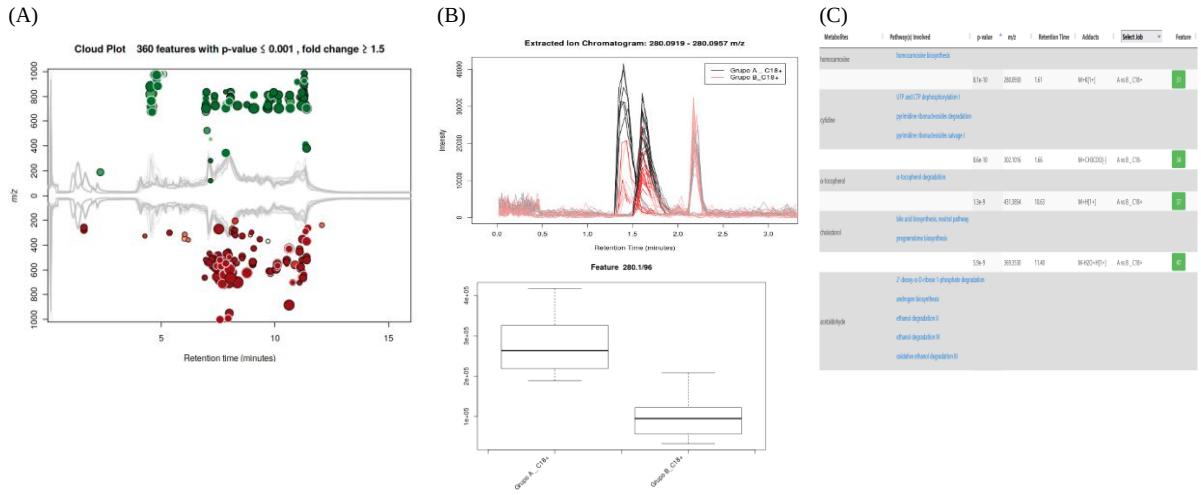
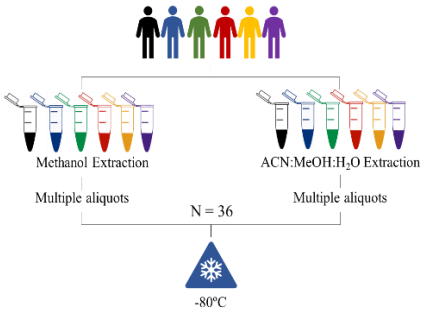


Figure 3.3.3 - Examples of data collected from XCMS online. (A) Cloud Plot generated from pairwise comparison of 2 distinct groups in the same LC-MS acquisition mode, in this case is group A vs group B (group explanation in chapter 3.4) in the same experimental setup. From the pairwise job both, a chromatogram alignment (B) as a dual boxplot comparison for a given *m/z* (treated as a feature) is retrieved to compare each group. Finally, a multimodal job (combines identified data by sample only, regardless of chromatographic or spectrometric acquisition conditions is executed, and the table retrieved is like shown in (C) there all features detected in the pairwise comparison are compared against the METLIN [114] human metabolome database for metabolite identification.

3.4. Patient Groups Assembly

Two sets of samples were analysed in this work. A first one compares the impact of the solvent extraction method (referenced earlier) on two groups of patients. A second assay compares 3 distinct groups to uncover differences between them with the analytic tests used. A schematization of these two assays is shown in the figures below (figure 3.4.1 and figure 3.4.2).



Group / Extraction	Group 1 (Deceased)	Group 2 (Survivor)
Methanol	Group A	Group C
ACN:MeOH:H_2O	Group B	Group D

Figure 3.4.1-Representative scheme of patient and sample logistics. 6 patients were chosen in accord with variables shown in this chapter, and each sample was processed with the two extraction methods, separately. Each extracted sample was divided into three aliquots that were stored in -80 ºC till FTIR and MS analysis, totalizing 36 samples. Samples were further identified as GROUP A, B, C and D according to extraction submitted and according to the patient's outcome later explained.

Patients for the first assay were selected and grouped according to the following criteria (figure 3.4.1):

- Group 1 – Male deceased COVID patients who needed ECMO, with ages between 50 and 70 years old
- Group 2 – Male COVID patients who needed ECMO, with ages between 50 and 70 years old

Take note that samples from each group (1 and 2) were processed with two extraction methods, methanol, and Acetonitrile:Methanol:Water (ACN:MeOH:H_2O). Samples from group 1, extracted with methanol, will be called group A, and those extracted with ACN:MeOH:H_2O will be called group B. Samples from Group 2 submitted to methanol and ACN:MeOH:H_2O will be called group C and group D, respectively. More information about the patients' characteristics is given in Table 3.4.1.

Table 3.4.1 - Patients' demographics. Due to sample size, each variable is expressed for each patient to facilitate assessment. (Abbreviations: BMI: Body Mass Index; ECMO: Extracorporeal Membrane Oxygenation; ICU: Intensive Care Unit; PCR: Polymerase Chain Reaction)

		Survived (n = 3)	Deceased (n=3)
Variables			
Age; years		56; 56; 54	52; 58; 63
Body/Mass Values	Weight; kg	80; 110; 75	80; 150; 80
	Height; cm	-; 175; 180	175; 170; 180
	BMI; kg/m^2	-; 35.92; 23.14	26.12; 51.9; 24.69
Hospital Variables	Arterial Hypertension (N)	3	2
	Obesity (N)	1	1
	Diabetes Mellitus (N)	1	2
	Dyslipidaemia	0	2
	Asthma (N)	0	1
	Days with Mechanic Ventilation; days	4; 23; 74	14; 16; 4
	Days with ECMO; days	9; 6; 48	3; 10; 3
	Days in ICU; days	41; 29; 86	14; 16; 4
	Days between symptoms onset/PCR and ICU admission	13; 8; 10	12; 5; 10
Lab. analysis	Days between sample collection and patient decease	-	13; 9; 3

The second assay was performed with a new set of patient samples. Based on the results obtained with samples from the first set only the ACN:MeOH:H_2O extraction was performed (figure 3.4.2). The comparison groups were decided by the followed variables of choice:

- Group A – COVID-19 male patients who did not require ventilation.
- Group B - COVID-19 male patients who went through ventilation.
- Group C - COVID-19 male patients who went through ventilation and died in ICU.

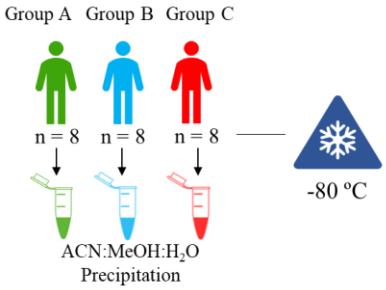


Figure 3.4.2 - Representative scheme of patient and sample logistics. 24 patients were chosen in accord with variables shown in this chapter. Samples were processed using the ACN:MeOH:H_2O method. The total 24 resultant aliquots were stored at -80ºC.

Table 3.4.2 - Patients' demographics. (Abbreviatures: BMI: Body Mass Index; IQBV- Inter-Quartile Boundary Values [Percentile 25; Percentile 75]; COPD- Chronic Obstructive Pulmonary Disease; OSA- Obstructive Sleep Apnoea; ECMO: Extracorporeal Membrane Oxygenation).

		Group A (n = 8)	Group B (n = 8)	Group C (n = 8)
Variables				
Body/Mass Values	Age, years, (median; IQBV)	52.00 (43.25; 62)	55.00 (51; 66.75)	65.00 (60.25; 67.75)
	Weight, kg, (median; IQBV)	80.00 (72,50; 100.00)	95.00 (80.00; 117.50)	90.00 (80.00; 100.00)
	Height, cm, (median; IQBV)	170.00 (167.50; 180.00)	170.00 (170.00; 187.50)	175.00 (175.00; 175.00)
	BMI, kg/m^2, (median; IQBV)	27.68 (26.91; 27.68)	31,18 (24.44; 35.66)	29,39 (26.12; 33.41)
Number of days in ICU, days, (median; IQBV)		6.20 (4.65; 8.48)	13.42 (10.27; 21.36)	11.65 (7.37; 14.60)
Comorbidities, n		7	7	5
Comorbidities, %	Arterial Hypertension	25	75	62.5
	Diabetes Mellitus	25	37.5	37.5
	Dyslipidemia	25	25	37.5
	Obesity	37.5	62.5	25
	Asthma	25	25	0
	COPD	12.5	0	25
	OSA	0	12.5	25
	Alcoholism	0	0	12.5
	Smoking	0	0	25
	Atrial Fibrillation	0	0	12.5
	Depression	12.5	0	0
	Peripheral arterial disease	12.5	0	0
	Hyperuricemia	0	0	12.5
	Stroke	0	0	12.5
	Carcinoma	0	0	0
ECMO, n		0	0	12.5
Lab. Analysis	Days between sample collection and patient decease, (median; IQBV)	-	-	5,00 (4.25; 11,50)

Chapter 4 – Results & Discussion

4.1. FTIR Analysis Results

4.1.1. First Assay Results

The first step of this work involved testing two different metabolite extraction protocols, one using methanol and the other using a mixture of acetonitrile:methanol:water, as laid at in spectra below. Results from a small set of patients (figure 4.1.1.1) are presented and described below. Figure 4.1.1.1 shows the results obtained when testing the atmospherically corrected spectra versus no pre-processing.

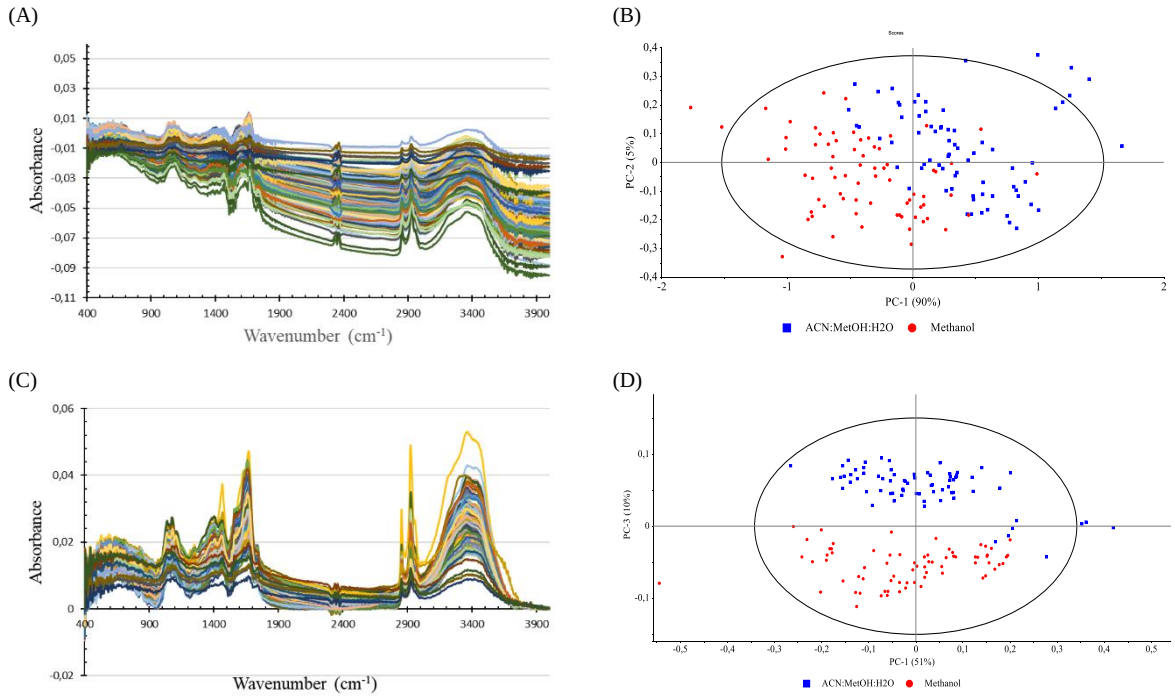


Figure 4.1.1.1 – (A) Spectral data retrieved from FTIR analysis, corresponding to Raw data (no pre-processing). (C) Spectral data when performed Atmospheric Correction. (B) Scores diagram from data (A) with Hotelling's T2 ellipse at 5% significance for PC1 (90% variance) versus PC2 (5% variance). (D) Scores diagram from data (C) with Hotelling's T2 ellipse at 5% significance for PC1 (50% variance) versus PC3 (10% variance). Program used, The Unscrambler X®

Based on observations, the prepossessed (figure 4.1.1.1, B) spectra allowed better separation between samples that were submitted to methanol extraction from samples submitted to ACN:MeOH:H_2O extraction.

On figure 4.1.1.1, D, PC3, with variance of 10%, had a major role in separating both extraction methods, since an imaginary line can be draw perpendicularly to the PC3 axis to separate both groups. For that reason, the PC3 loading score plot was extracted (figure 4.1.1.2).

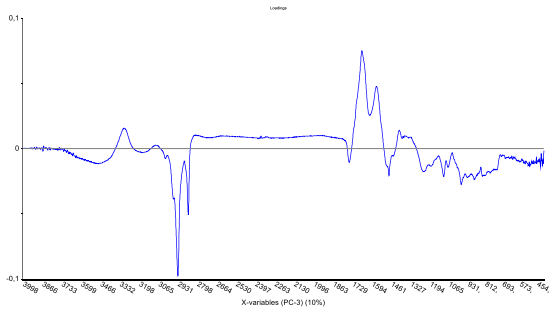


Figure 4.1.1.2 - Loading plot for PC3 from figure 4.1.1.1(D). Loading scores are always limited between 1 and -1, which are expressed on the Y axis of the plot. On the X axis, are the corresponded variables that the principal components use, which are the wavenumbers (from 400 to 4000) used for FTIR analysis.

The loadings from each principal component describe the contribution from the variables and their correlation to the PC. Every PC has all variables (wavenumber in the case of FTIR) analysed, in this case, and the loading value describes how well a PC considers a certain variable/wavenumber. If a variable describes completely a PC, then its value is positive and close to 1, and samples that are well explained by that variable will be placed on the positive side of the PC in the score plot. In the other hand, if a variable does not contribute to a PC the loading value for that variable is very close to zero, however, if it has a negative loading value for that variable, the PC is inversely correlated, meaning that samples that have more tendency to be explained by that negative variable loading value will be placed on the negative side of the principal component in the score plot. Variables with close to zero loading scores can be overlooked as they do not contribute to separation along that principal component.

And how does this loading score is interpreted in the PCA in the case of FTIR analysis? For a wavenumber with a positive high value on the loading score, all samples with a positive high score in that PC have higher values than average for that given variable. For a wavenumber with a loading value highly negative (closer to -1), the samples that have a negative score on that PC, in the PCA, have lower values than average. Meanwhile, samples with negative score on the PCA, for the PC in discussion, have a negative correlation with positive high values and a positive correlation with negative high values.

In this way, one can know which wavenumbers contribute to the separation in the PCA plot, and thus deduce what group bond, or groups of bonds, are mainly contributing for the separation of the samples and finally assume what main family of molecules each group has (e.g., lipids, amides, and phospholipid esters). As observed on figure 4.1.1.3, samples extracted with methanol have higher concentration/quantity of lipidic content, revealed by the loading peak of almost -0.1 on the wavenumbers 2924 cm^{-1} and 2853 cm^{-1}, corresponding to vibrational frequencies from $-(CH_2)_n$ groups from samples' lipidic content. On the other hand, the cluster formed by samples extracted with ACN:MeOH:H_2O had a higher influence from wavenumbers 1655 cm^{-1} and 1556 cm^{-1}, denoting contribution from variables representing the functional group from amides (O=C-N-H), revealing major contribution from peptide bonds for separation from samples extracted only with methanol.

It can also be stated that the influence of different solvents/mixtures used in extraction is being detected, since on one side, molecular vibrations associated with nonpolar molecules (lipids) can be observed, and on the other, amide groups from highly polar molecules.

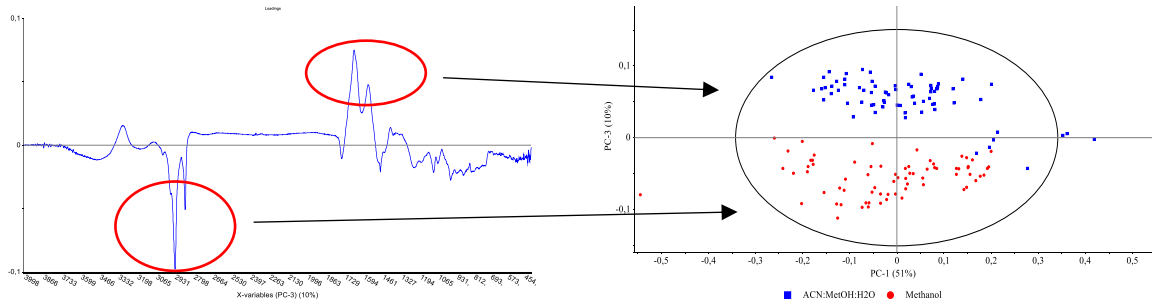


Figure 4.1.1.3 – Loading scores from PC3 enlightened in a red circle, majorly responsible for the separation of the two clusters on the PCA (right) from figure 4.1.1.1(C).

To minimize interferences and enhancing the biological/chemical signature from data, data pre-processing procedures announced on chapter 3.2 were performed and evaluated by PCA. Various spectra pre-processing techniques were evaluated by PCA, regarding the separation between samples that went thought the two distinct extraction methods, and the best results are shown on figure 4.1.1.4.

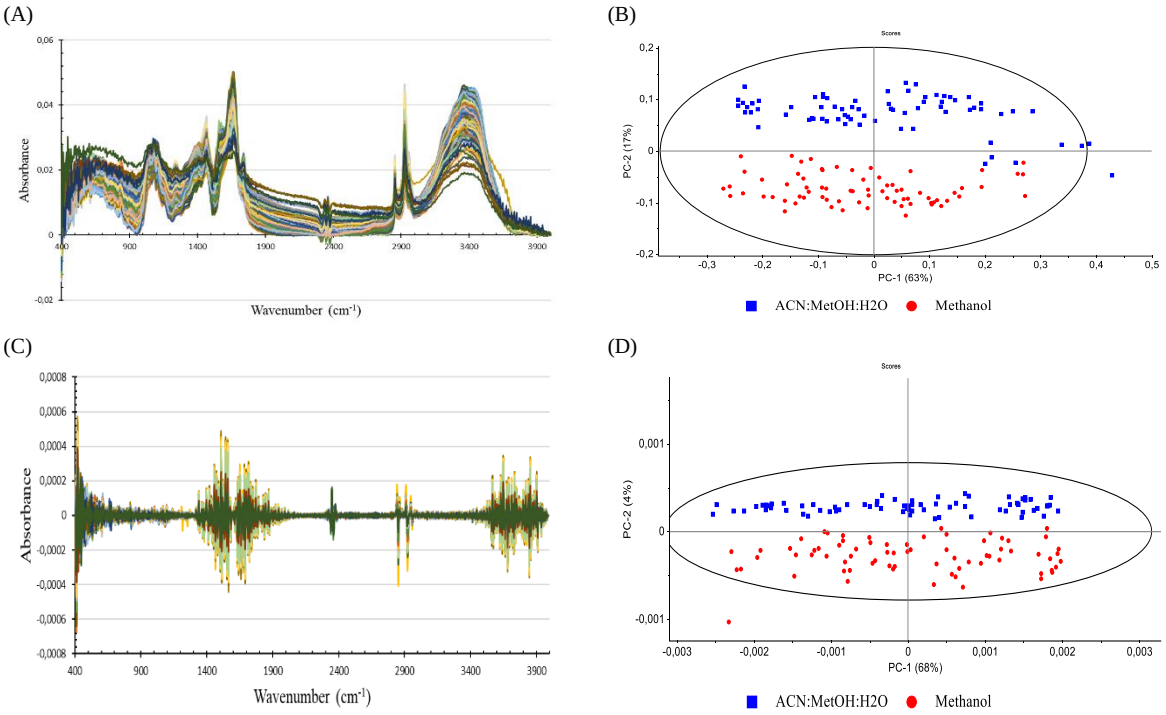


Figure 4.1.1.4 - (A) Spectra resultant from Atmospheric and Baseline correction, and Unit Vector Normalization. (B) PCA (PC1 63% variance against PC2 17% variance) correspondent to spectra (A). (C) Spectra resultant from Atmospheric correction and Savitzky-Golay filter with a 2nd order polynomial and a 15-points window. (D) PCA (PC1 with 68% variance against PC2 with 4% variance) resultant from spectral data (B).

For the best pre-processing techniques (ATM+BC, ATM+BC+UVN, ATM+SG 2nd derivate 15^{pnts} window), spectral outliers were identified (reported on table 4.1.1.1) according to the explained variance, F-residuals in function of Hotelling's' T2 and Hotelling's ellipse both at 5% significance. New analyses were conducted for the respective pre-processing without those outliers.

Table 4.1.1.1 - Samples identified as outliers in replicates of FTIR serum spectra from the 6 patients samples, extracted with methanol and ACN:MeOH:H2O extraction protocols, and submitted to the three best considered pre-processing corrections.

Pre-processing	Patient(s)	Extraction method	ID designation excluded
Atmospheric correction + Baseline correction	1	Methanol	1_4
	4	ACN:MeOH:H_2O	22_4
	5	ACN:MeOH:H_2O	29_2
	6	ACN:MeOH:H_2O	36_1; 36_2; 36_4
		Methanol	33_2
Atmospheric correction + Baseline correction + Unit Vector Normalization	6	Methanol	33_2
		ACN:MeOH:H_2O	35_4; 36_1; 36_2; 36_4
Atmospheric correction + SG (2nd derivate 15^{pnts} window)	1	Methanol	1_4
	3	ACN:MeOH:H_2O	16_4

Hierarchical Clustering Analysis (HCA) was attempted for the three pre-processing methods mentioned above. HCA for Unit vector Normalization applied to atmospheric and baseline correction data, achieved 77,8% accuracy for the outcome (death or survival), regardless of the extraction method. Remaining results for the other pre-processing methos did not reveal any type of clustering regarding the outcome of the patients (survival / deceased) or the method of extraction used for the samples. These analyses were applied for both all sample population and for samples without replicates (medians used). Result for the better result is shown on figure 4.1.1.5.

Misclassified patient samples from UVN+ATM+BC HCA are corresponded to patient 6 (misclassified as dead), patient 5 (misclassified as alive), and patient 4 (who had only one sample misclassified as dead). Possible causes for misclassification of patient 5 can be attributed to the fact that even though being the older patient from the group, with 63 years old, had the lowest BMI, no apparent associated comorbidities, with the lowest number of

days submitted to mechanical ventilation (4 days compared to 14 and 16 days from the other patients) and as well the lowest number of days in hospital (7 days compared with 17 and 23 from the patients who died). These variables could confuse patient 5 as possible candidate to survive, having this patient a biological signature more relatable to patients from the survival group. The specific outcome of this patient was clinical picture aggravation concerning pulmonary, and pulmonary tract atelectasis (total to partial collapse), and possible inflammatory exacerbation due to bacterial infection acquisition (non-identified) on the upper respiratory tract. Due to complexity and irreversibility of this patient condition, patient 5 obituary occurred upon ECMO suspension.

Patient 6, being part of the alive group by the time of the analysis (outcome unknown due to hospital transference), is probably misclassified as a dead patient derived from the fact that this patient had a serious respiratory insufficiency with pulmonary superinfection of *MSSA* and *K. pneumoniae* cumulatively with fungal infection with *Candida albicans*. Adding a myopathy due to prolongated ECMO and ventilatory support, this patient must have had a serum biological signature resembling a possible death outcome when compared with patient 6.

These findings should be compared to a broader patient database in order to determine and weight these conclusions, which may be skewed or biased owing to a small sample size analysis. A simple answer to this problem is to continue this research within the context of the PREMO project, which has collected over 4000 samples from over 350 patients so far.

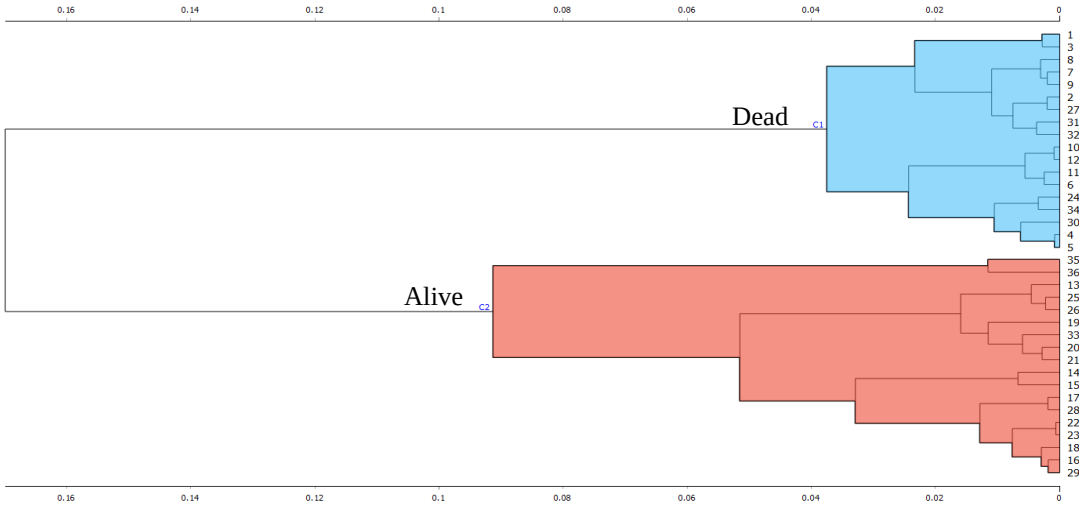


Figure 4.1.1.5 - Hierarchical Cluster Analysis for serum pre-processed with atmospheric and baseline correction with unit vector normalization. HCA complete linkage method and cosine correlation distance metric was used. Medians for replicas were used for this analysis. 8/36 samples were misclassified, corresponding to samples 31, 32 and 34 (corresponding to extraction samples from patient 6 submitted to methanol extraction); sample 24 (sample from patient 4 from ACN:MeOH:H_2O); samples from patient 5 (sample number 28 and 29 from ACN:MeOH:H_2O, and 25 and 26 from methanol extraction). Program used, Orange 3.32.

Machine learning techniques were applied to the 3 pre-processed databases, to infer the data quality to predict the patients' outcome. Results are shown in the table below (table 4.1.1.2). All data was used to conduce this analysis, and as for the parameters used in the test and scores, they were done using Leave One Out (LOO).

In order to not entering too deep in the explanation of each algorithm used, a brief description of each machine learning model will be given. Support-vector machine (SVM) is a learner that tries to maximizes the data space between the attributes (class variables of interest), given the name, hyperplane, and in this way maximizes the margin between instances of different classes [118]. Neural Network uses a multi-layer perceptron algorithm, making this model appropriate for learning linear and non-linear data models [119]. AdaBoost (from Adaptive Boosting), the principle behind boosting algorithms is first we built a model on the training dataset, then a second model is built to rectify the errors present in the first model. This procedure is continued until and unless the errors are minimized, and the dataset is predicted correctly[120]. k-nearest neighbours (kNN) algorithm consists in data classification according to the proximity of the test data, this model relies on the distance for classification [121]. Stochastic gradient descent (SGD) evaluates how well the model fits the data by finding the sum of the squared

residuals' derivative lowest value that corresponds to the model that fits better to all variables in the data, but this process is too slow for data with hundreds of variables, so this model uses a stochastic (random) algorithm that accelerates this process without compromising the results [122].

Table 4.1.1.2 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for target (death / survival): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall).

	ATM+BC					ATM+BC+UVN					ATM+S-Golay				
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**AUC**	**CA**	**F1**	**Precision**	**Recall**
SVM	0.97	0.93	0.93	0.94	0.93	0.98	0.95	0.95	0.95	0.95	0.73	0.68	0.68	0.68	0.68
Neural Network	0.99	0.97	0.97	0.97	0.97	0.98	0.97	0.97	0.97	0.97	0.84	0.77	0.77	0.77	0.77
AdaBoost	0.74	0.74	0.74	0.74	0.74	0.92	0.92	0.92	0.92	0.92	0.81	0.81	0.81	0.81	0.81
kNN	0.89	0.82	0.82	0.82	0.82	0.96	0.92	0.92	0.92	0.92	0.44	0.48	0.48	0.48	0.48
SGD	0.94	0.94	0.94	0.94	0.94	0.98	0.98	0.98	0.98	0.98	0.83	0.83	0.83	0.83	0.83

Evaluation of the 5 predictive models on this set of wavenumbers (400 to 4000 cm^{-1}) in the 3 pre-processed data sets showed best results using Stochastic Gradient Descend regression model on data pre-processed with atmospheric and baseline correction scaled with unit vector normalization, therefore all following results are based on this model and dataset . The average area under the curve (AUC) was calculated at 0.98 for outcome (death or survival), demonstrating good predictive power of serum FTIR analysis. Representative receiver operating curve (ROC) of one model from one pre-processed data is shown in figure 4.1.1.6.

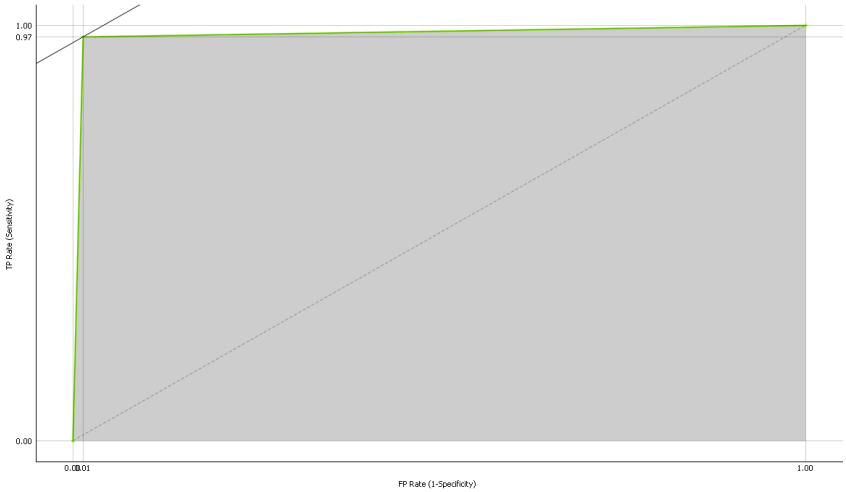


Figure 4.1.1.6 - Predictive model based on the 400 to 4000 cm^{-1} wavenumbers for patient outcome. 144 samples total ran through the Stochastic Gradient Descent model, ran on default parameters on Orange 3.32.

Confusion matrix from the regression model SGD is shown below (table 4.1.1.3)

Table 4.1.1.3 – Confusion matrix for patient outcome, from SGD regression model. Percentages relative to proportion of predicted samples.

		Predicted		
	Outcome	Death	Survival	Σ
Actual	Death	97.3%	1.4%	72
	Survival	2.7%	98.6%	72
	Σ	73	71	144

Only three of the 144 samples were incorrectly categorised. Two samples from patient 3 were incorrectly identified as dead, whereas just one sample from patient 5 was incorrectly identified as living. Two of these three samples were extracted using the methanol extraction methodology (one from patient 3 and one from patient 5), with the left sample from patient 3 being extracted with the ACN:MeOH:H2O protocol.

Workflow carried can be observed below (figure 4.1.1.7).

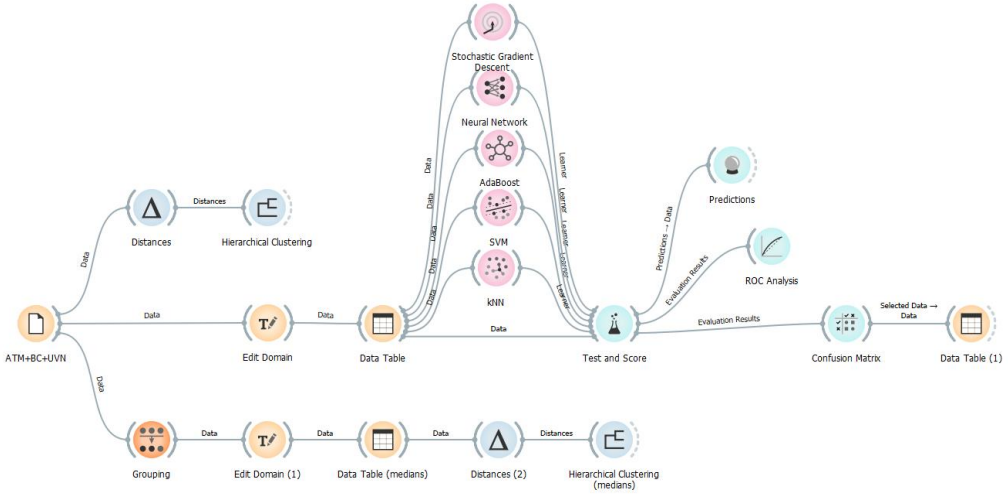


Figure 4.1.1.7 – Workflow for prediction using various machine learning algorithms. HCA analysis was conducted for all replicates (144 samples), and for median values (36 samples). This workflow ran three times, corresponded to the three data pre-processing techniques chosen. Consider that analysis took over one hour for each pre-processing due training and test methods adopted.

With this chapter it can be concluded that, regardless of the method of extraction used, FTIR analysis is capable of distinguishing dead patient samples from alive ones. One limitation associated with this work is that the number of patients is diminished and are not enough to create a robust enough model to be put in practice on the real world. A solution would be through continue this work and enlarge the dataset so it could be used as a database, and to be tested with new samples to corroborate the results obtained.

4.1.2. Second Assay Results

As this work has the objective to focus mainly on metabolomics and not so much in the lipidomics field, the ACN:MeOH:H2O extraction was selected and applied to the 2nd set of patients to identify any differences in the spectra related to the outcome (death or survival), and condition in ICU care (invasive ventilation or no ventilation). 24 patients FTIR data was a PCA analysis on The Unscrambler X. Cluster from the 3 different groups formed by these patients was not revealed at first sight, however, it was identified a possible sample outlier due to the fact of having a totally different scale compared to the remaining samples. As it was only one replicate (sample 40 replicate number 2), it was excluded from further analyses (figure 4.1.2.1).

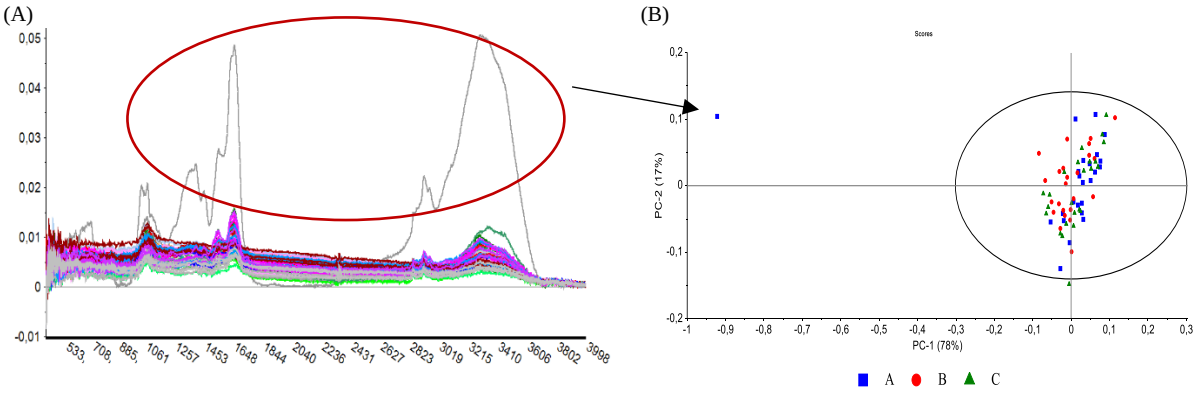


Figure 4.1.2.1 – (A) Spectra from data pre-processed with atmospheric and baseline correction from all patients of second study assay. (B) PCA from spectra mentioned, PC1 with 78% variance against PC2 with 17% variance.

Removing the second replicant from patient 40, permitted ameliorate the explained variance from the atmospheric and baseline correction data PCA (figure 4.1.2.2)

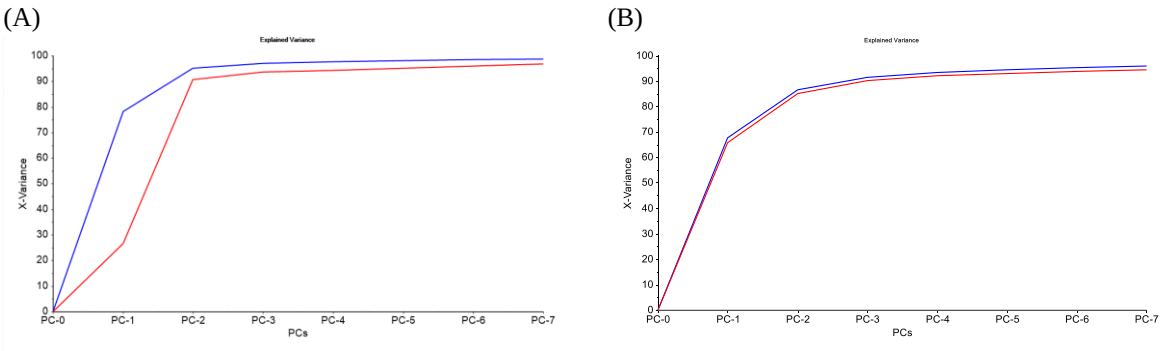


Figure 4.1.2.2 – Explained variance for PCA (A) before patient 40 second replicate removal and (B) after removal. (blue line: calibration; red line: validation).

To find differences between sample groups, pre-processing was performed, namely, Unit Vector Normalization and Savitzky-Golay filter. The PCA with best separation between classes is demonstrated on figure 4.1.2.3. Maximum Normalization was applied to atmospheric and baseline corrected data, retrieving among the pre-processing methods used, the best clustering for the COVID positive patients who went through ventilation, who went through ventilation and died and survivors who did not need ventilation. It also had the higher explained variation compared to the other principal components from the other pre-processing PCA data. For these reasons following results are based on the atmospheric and baseline correction followed by maximum normalization model.

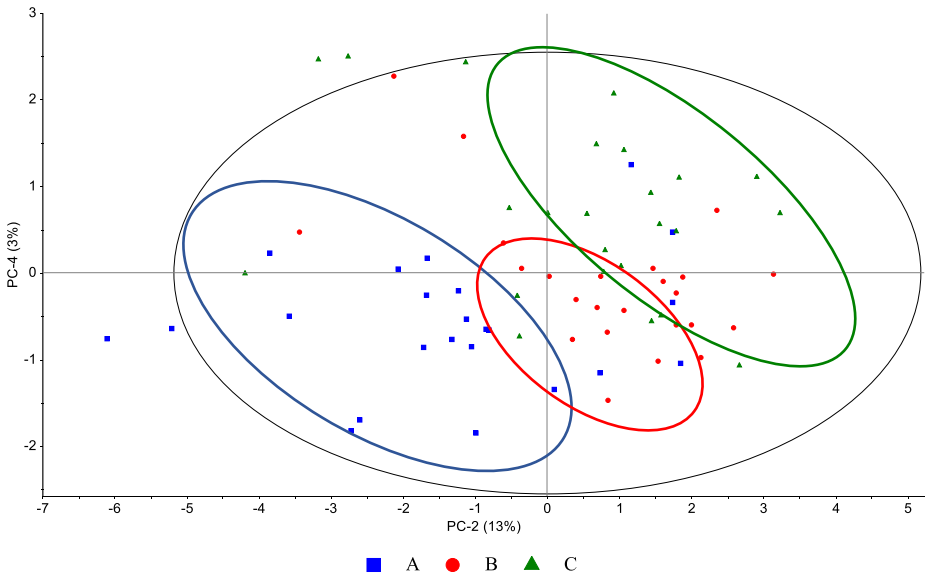


Figure 4.1.2.3 – PCA score (PC2 13% variance against PC4 3% variance) from dataset pr-processed with atmospheric and baseline correction followed by maximum normalization, which divides each absorbance from a sample by its maximum value. All samples with all replicates from all groups were utilised apart from sample from patient 40 replicate number 2.

Loading score for PC2 from figure 4.1.2.3 showed high positive values for wavenumbers ~3346 cm^{-1} to ~3286 cm^{-1} and for wavenumbers ~1655cm^{-1}, giving indication that sample separation occurred mainly by the influence, on PC2, of N-H bond stretching plausible from amides present in peptides or even RNA/DNA, possible implying a higher inflammatory state from group to group, being the non-ventilated alive group (group A) the patients with the least amount of free circulating peptide in their serum and deceased patients submitted to invasive mechanical ventilation (group C) with the higher amount of these peptide molecules in their circulation, indicative of a higher immune system response [123]. These findings are similar to those reported by Zhang *et al* [124] that concluded that that severity and disease outcome from SARS-CoV-2 infection was predicted from higher inflammatory associated peptides.

In the same loading score mentioned above a major noise signal was detected below 800 cm^{-1}. Purging these data columns was done to ameliorate the PCA loading and score (figure 4.1.2.4).

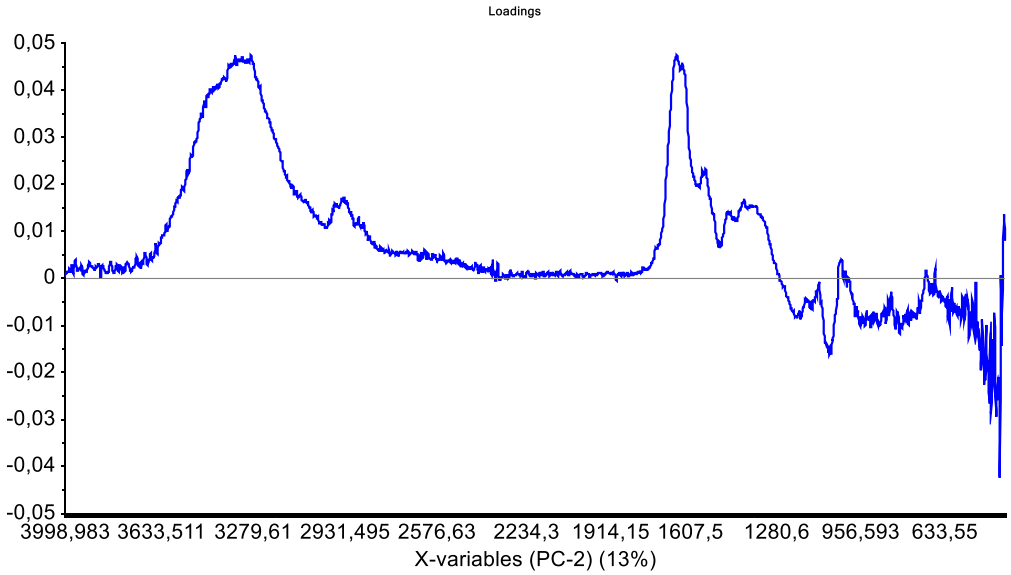


Figure 4.1.2.4 – Loading Score from PC2 from figure 4.1.2.3. As observed this principal component has a large loading weight >0.4 from wavenumber ~3346 to 3226 cm^{-1} and another peak at ~1655 cm^{-1}. From the 800 cm^{-1} below the loading presents heavy negative values despite of the PCA report saying that this band was not impacting the Score.

As expected, removal of data between 800 cm^{-1} and 400 cm^{-1} enhanced the performance of the PCA. The variation for PC2 increased from 13% to 16%, and explained the variance increased from 61,7% to 70,04% validation, starting from PC1.

The next step was to select sample outliers according to the PCA Hoteling's T^2 at 5% significance and F-residuals on influence graphic, observed on table 4.1.2.1.

Table 4.1.2.1 – Outlier identification. From dataset pre-processed with atmospheric and baseline correction followed by maximum normalization.

Outliers			
Patient nr	Group	Replicate	ID excluded
21	B	2	21_2
147	A	1 ; 2	147_1; 147_2
191	A	2	191_2

As to understand better the separation between the three sample groups, comparation in pairs was used as followed: COVID-19 positive patients who did not need ventilation against those who went through ventilation (figure 4.2.1.5); COVID-19 positive patients who survived without ventilation against those who died and went through ventilation (figure 4.1.2.6); finally, COVID-19 positive patients that survived against those who died, both groups went through ventilation (figure 4.1.2.7).

For patients who survived, and comparing the ventilation factor, PCA (figure 4.1.2.5) shows a tendency for separation along the principal component 2, with the samples from patients who did not receive ventilation being preferably allocated on the 1st and 4th quadrant of the PCA, and the patients who went through ventilation on the 2nd and 3rd quadrants. This separation along PC2 has a major contribution from the wavenumbers ~3570 cm^{-1} to ~3000 cm^{-1} and ~1660 ~1630 cm^{-1}. Outlier samples were excluded during the PCA and testing classification algorithms (ID samples considered as out of cluster: 427_3; 401_1; 401_2; 21_2; 191:2; 191_3; 110_1) remaining 40 samples in total.

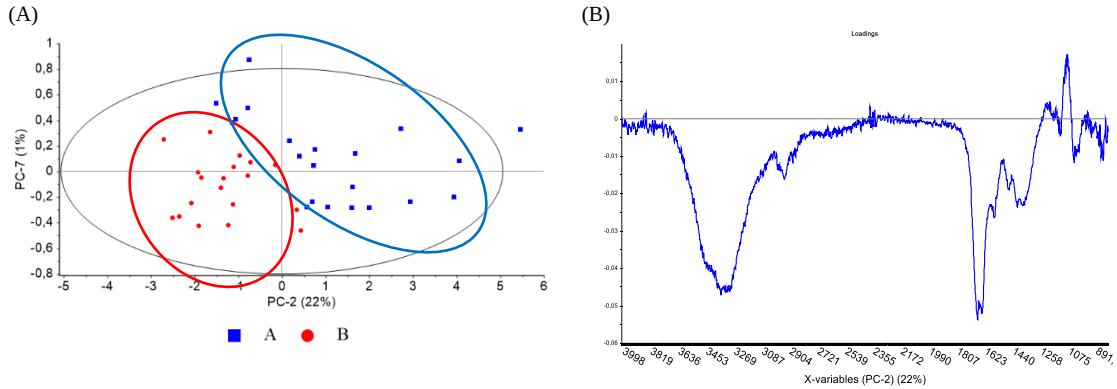


Figure 4.1.2.5 – (A) PCA (Pc2 against PC7) between samples from patients who did not go through ventilation (Group A) and those who went through ventilation (Group B), both survived. (B) Principal component 2 loading score, the main PC that separates group A and B. It can be observed the main wavenumbers that influence this PC.

Hierarchical clustering analysis did not reveal significant information regarding the condition of each patient. Machine learning algorithms were applied to the data, and the same algorithms as the chapter were employed, results are shown on table 4.1.2.1.

Table 4.1.2.2 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for target (Ventilation): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall).

	AUC	**CA**	**F1**	**Precision**	**Recall**
SVM	0.89	0.80	0.80	0.80	0.80
Neural Network	0.84	0.85	0.85	0.87	0.85
AdaBoost	0.72	0.72	0.72	0.73	0.73
kNN	0.86	0.75	0.75	0.75	0.75
SGD	0.81	0.80	0.80	0.81	0.80

Table 4.1.2.3 - Confusion matrix for patient condition: Ventilation status, from Neural Network prediction model. Percentages relative to proportion of predicted samples.

		Predicted		
	Condition	Non-Ventilated	Ventilated	Σ
Actual	Non-ventilated	94.1%	21.7%	21
	Ventilated	5.9%	78.3%	19
	Σ	17	23	40

Patients 40 (only one replicate), 191 (2 outliers samples and last as misclassified), 456 (only one replicate) , 13 (two replicates) were misclassified as ventilated subjects. Only one replicate from patient 28 was misclassified as non-ventilated.

For patients who had all their samples classified as outliers and misclassified, as patient 191, possible cause of misclassification could be due to variability in the demographics variables. This patient is one of the two obese patients of the group where it's inserted with no more comorbidities to point, and accordingly with Huang *et al* a patient solely diagnosed with obesity has a odds ratio of 2.19 for invasive mechanical ventilation (IVM) requirement [125]. More serum from individuals with this comorbidity should be collected and analysed using FTIR to prove this claim, as only two obese people are insufficient to statistically validate this statement.

Patient 13's major comorbidity is bronchial asthma, and even though this patient did not undergo ventilation, this comorbidity alone might trigger a metabolic shift, resulting in a blood biosignature similar to those in the ventilation group, since asthmatic patients have a higher tendency for IVM and COVID-19 severe state [126]. The

same statement written before is applied for this comorbidity, more samples from patients with this condition need to be gathered to validate this conclusion

Between non-ventilated and ventilated patients, when death variable is analysed in PCA (figure 4.1.2.6), the principal component 2 persists as the main component to split alive patient samples from dead patient samples. Alive patients tend to be more dispersed and localized on 2nd and 3rd quadrant while ventilated death patient samples tend to be preferably assigned on 1st and 4th quadrant. Wavenumbers ~3370 cm^{-1} to 3295 cm^{-1} and ~1656 cm^{-1} continue to have influence on PC2, and in this case samples from dead ventilated patients have higher than average values of absorbance for said wavenumbers, suggesting that these samples have a higher concentration of molecules correspondent to these chemical bonds than samples from alive patients.

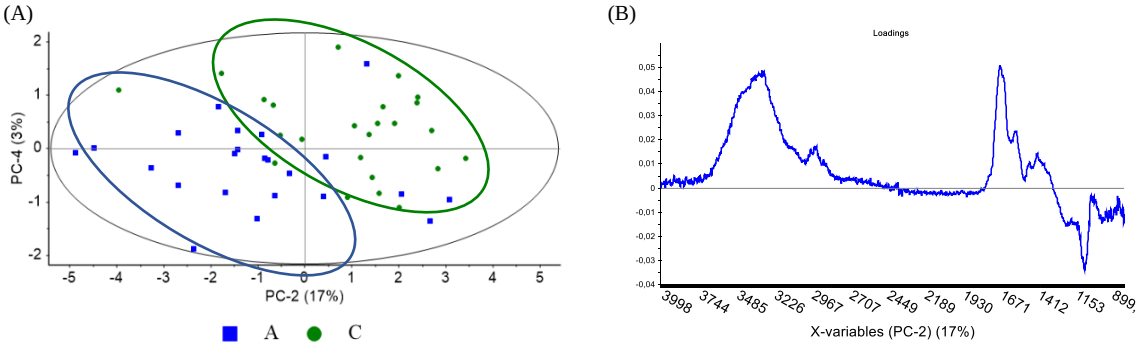


Figure 4.1.2.6 - A) PCA (PC2, 17% variance against PC4, 3% variance) between samples from patients who did not go through ventilation and survived (Group A) and those who went through ventilation and died (Group C). (B) Principal component 2 loading score, the main PC that separates group A and C. It can be observed the main wavenumbers that influence this PC.

Cluster analysis, as predicted, did not retrieve any kind of relationship between the samples of the same group. Various combinations of type of linkage and distance metric were modified get a better solution. However, in this real-world application, this data is denoted with higher dimensions and possibly not best suitable to be represented on a plot like this, which implies that poor correlations may be discovered, and worse, that this poor solution might not be identified as such by this type of algorithms.

Due to the small patient group, the Leave One Out (LOO) validation was utilised to assess the learners, and the findings are shown in table 4.1.2.3.

Table 4.1.2.4 - Test and scores result for all learning method. Sampling type: Leave One Out (LOO); Evaluation results for outcome, between samples from non-ventilated alive patients (group A) and ventilated deceased patient samples (group C): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall).

	AUC	**CA**	**F1**	**Precision**	**Recall**
SVM	0.84	0.77	0.77	0.77	0.77
Neural Network	0.93	0.85	0.85	0.85	0.85
AdaBoost	0.57	0.57	0.57	0.57	0.57
kNN	0.86	0.77	0.77	0.77	0.77
SGD	0.85	0.85	0.85	0.86	0.85

The best learner was Neural Network with overall best area under the curve, precision, and recall, confusion matrix for the given learner is given on table 4.1.2.4.

Table 4.1.2.5 - Confusion matrix for patient outcome: death, from Neural Network prediction model. Percentages relative to proportion of predicted samples.

		Predicted		
	Outcome	Survival	Death	Σ
Actual	Survival	86.4%	16%	23
	Death	13.6%	84%	24
	Σ	22	25	47

Samples misclassified by this model were on replicate from patient 40, which can be seen on the respective PCA well defined in the wrong cluster, being the remain sample well classified as alive. Patients 31, 37 and 84 had their 1st replicate wrongly classified as alive patients samples, the remaining replicates from the respective patients were all well classified.

Patient 13 had all replicates wrongly classified on the deceased patients' group, with the same result reported on table 4.1.2.2. This patient with bronchial asthma could well be a candidate for ventilation. A comparation of samples along this patient internment during ICU stay is advised to denote any FTIR biosignature shift and compared again the different samples against non-ventilated, ventilated, and deceased ventilated patients' samples.

Ventilated alive patient samples and ventilated deceased patient samples were compared and the PCA that show relevant differences between these two groups is shown below (figure 4.1.2.7).

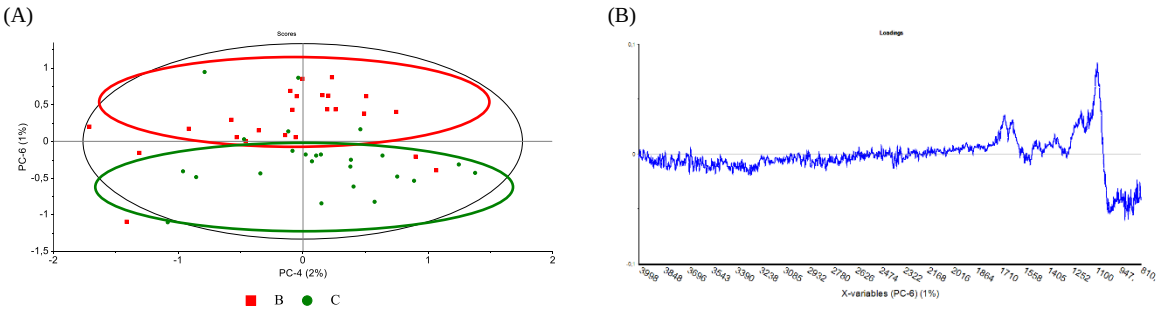


Figure 4.1.2.7 - A) PCA (PC4 against PC6) between samples from patients who went through ventilation and survived (Group B) and those who went through ventilation and died (Group C). (B) Principal component 6 loading score, the main PC that separates group B and C. It can be observed the main wavenumbers that influence this PC.

Machine learning algorithms scores are shown in the table below

Table 4.1.2.6 - Test and scores results for all learning methods. Sampling type: Leave One Out (LOO); Evaluation results for outcome, between samples from ventilated alive patients (group B) and ventilated deceased patient samples (group C): average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall).

	AUC	**CA**	**F1**	**Precision**	**Recall**
SVM	0.73	0.67	0.67	0.67	0.67
Neural Network	0.79	0.71	0.71	0.71	0.71
AdaBoost	0.69	0.69	0.69	0.69	0.69
kNN	0.80	0.77	0.77	0.77	0.77
SGD	0.69	0.69	0.69	0.69	0.69

kNN revealed to be the best learner, the respective confusion matrix is shown on table 4.1.2.6.

Table 4.1.2.7 - Confusion matrix for patient outcome and condition: death and ventilation status, from kNN prediction model. Percentages relative to proportion of predicted samples.

		Predicted		
	Outcome	Survival	Death	Σ
Actual	Survival	86.4%	16%	23
	Death	13.6%	84%	24
	Σ	22	25	47

Misclassified samples by the learning algorithm were from patients 427 (only one misclassified replica), 401 (one replica), 28 (one replica) and 110 (all replicas), misclassified as ventilated deceased samples. No apparent reason for patient 110 to be misclassified on the deceased patients' group was found, with 54 years on the day of the ICU admission, only ten days allocated on ICU facilities, 5 of which submitted to IMV. Among the group this patient is the one with less comorbidities, arterial hypertension.

Misclassified as alive ventilated patients, patients 79, 37 and 101 had one replica misclassified by the prediction model, patient 149 had 2 replicas misclassified. Patient 149, 67-year-old on day of ICU admission, stayed on this facility for 15 days, transferred to nursery on the sixteenth day, submitted to IMV for 16 days with no reported comorbidities. One of the main reasons for this patient to be misclassified as alive ventilated patient is that sample collection was done only two days after entry in ICU care, not giving enough time to settle the "right" biosignature denoted of a possible death outcome. A solution would pass through utilising, for example, a sample from this patient 24 hours before his death and see if the learning algorithm would still classify the sample as alive.

The workflow applied to this analysis was as followed on figure 4.1.2.8.

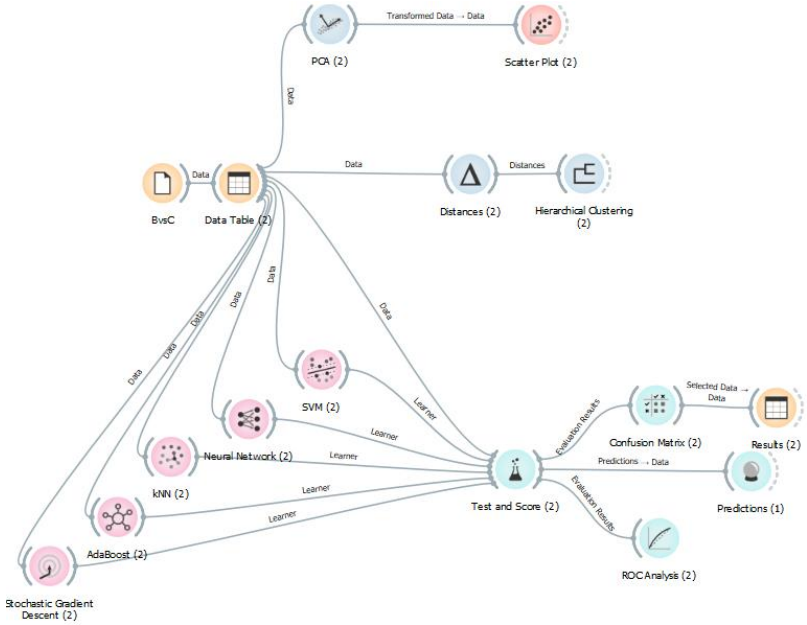


Figure 4.1.2.8 – Workflow performed on the three analyses from Chapter 4.1.2.

Overall sample classification from machine learning algorithms were good. Stands the same obstacle reported on the former chapter, the number of individuals that are studied are low. Noticeable separation from groups stands on the different expression level of N-H and C=O bonds (remarked from 3350 cm^{-1}, 3295 cm^{-1}, and 1650 cm^{-1}), which can be explained for a higher immune response state and a higher infection/inflammation response [124,127,128].

4.2. UHPLC-MS Analysis Results

4.2.1. First Assay Results – Extraction Procedure Comparison

The same methodology was applied as chapter 4.1.1, and the two extraction methods were compared. Data from untargeted MS-based metabolomics were analysed in a meta-analysis perspective, and the main metabolites found to be significantly changed between samples are showed below. Results retrieved from Group A versus B, for comparison of different extraction procedures, are given on table 4.2.1.1 and figure 4.2.1.1.

Comparing procedures for dead individuals yielded seven distinct metabolites from the chemical families of carboxylic acids (n=1), ketone bodies (n=1), vitamins (n=1), nucleoside compounds (n=1), lipids (n=2), and benzopyrans (n=1). The notable difference between the two methods was that all detected molecules had energy intensities greater for the methanol extraction but more substantial for the lipid molecule class, with a fold-change of roughly nine.

Table 4.2.1.1 - Metabolites Results from subgroup A vs B. (Abbreviations: R.T: retention time; F.C.: fold-change). Dysregulation in this context means how much higher the intensities of the metabolite of group B is compared to group A (e.g., α-tocopherol has a 3.99 fold-change with dysregulation "Down", this mean that group A has an intensity 3.99 times higher than B for the given molecule, and so, B is down-regulated relative to A).

Metabolites	*p*-value	*m/z*	R.T. (min)	Pairwise job	F.C.	Dysregulation
α-tocopherol	1.3e-9	431.3854	10.63	A vs B _ C18+	3.99	DOWN
(R)-lactate	1.2e-8	89.0247	2.10	A vs B _ HILIC-	2.1	DOWN
Cytidine	1.4e-7	242.0790	1.66	A vs B _ C18-	2.35	DOWN
4α-OCH_3-4β-CH-5α-zymosterol	2.8e-5	429.3748	10.63	A vs B _ C18+	9.02	DOWN
α-carboxyethylhydroxychroman	4.3e-4	279.1602	8.22	A vs B _ C18+	2.36	DOWN
(R)-3-hydroxybutanoate	1.1e-3	103.0402	2.07	A vs B _ HILIC-	1.59	DOWN

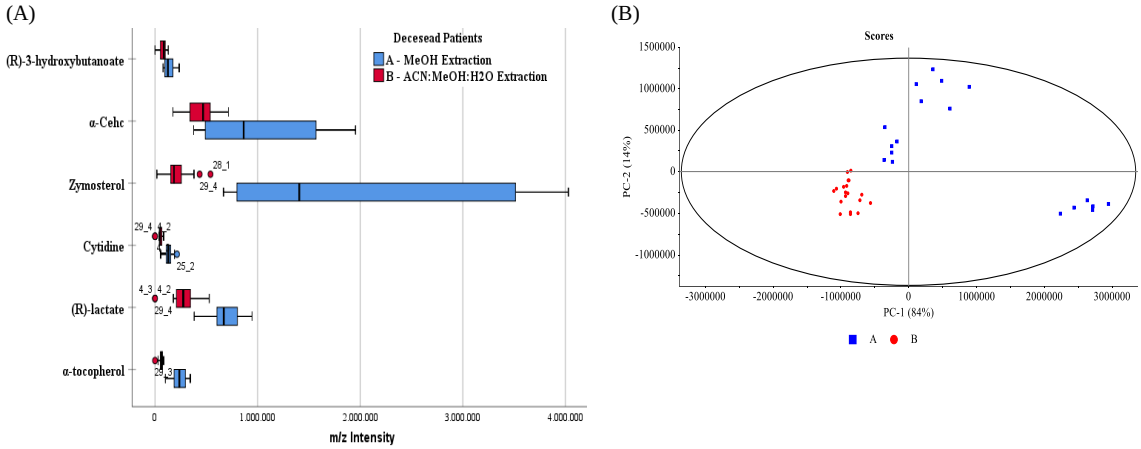


Figure 4.2.1.1 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.1. (α-Cehc is a synonym of α-carboxyethylhydroxychroman). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA resultant from allocating the m/z intensity of each metabolite of each group (A: Blue; B: Red) in The Unscrambler® X, is possible to see a fine discernment between group A and group B, being B better clustered in relation to A.

With these results is easy to establish that there is a clear distinction between the two methods of extraction (Methanol vs ACN:MeOH:H_2O). Although intensities are higher when looking to samples extracted with methanol, they have a greater dispersion as seen in figure 4.2.1.1. On the other hand, samples extracted with the ACN:MeOH:H_2O solvent mixture show higher reproductible data, and cluster better in the PCA analysis.

Next data continues to show the comparison between extraction methods but now among the samples of deceased patients, achieved by comparison of Group C and D (Table 4.2.1.2). Seven metabolites significantly dysregulated between each protocol were identified, belonging to the family carbohydrates and carbohydrates conjugates (n=1), hydroxy acids and derivates (n=2), lipids and lipid-like molecules (n=1), purine nucleotides (n=1), and pyridine / pyrimidine molecules (n=2). Lipid-like molecules continue to be down-regulated according

to results and concordant with the findings of the first comparison between protocols. Family classes that were retrieved with higher energy intensities (and extrapolating), in higher quantity were carbohydrates and carbohydrate conjugates, purine nucleotides, pyridines and derivatives, with a fold-change ranging between 1.97 to 16.31.

Table 4.2.1.2 - Metabolites Results from subgroup C vs D. (Abbreviations: R.T: retention time; F.C.: fold-change).

Metabolites	*p*-value	*m/z*	R.T. (min)	Pairwise job	F.C.	Dysregulation
D-sorbitol	3.2e-6	183.0854	1.68	C vs D _ C18+	1.97	UP
4-(3-pyridyl)-butanoate	4.4e-5	166.0848	2.27	C vs D _ C18+	16.31	UP
3'-dephospho-CoA	9.0e-5	688.1577	4.54	C vs D _ C18+	8.86	UP
(R)-3-hydroxybutanoate	4.4e-4	103.0405	2.09	C vs D _ HILIC-	1.78	DOWN
(R)-lactate	2.4e-2	89.0247	2.08	C vs D _ HILIC-	1.47	DOWN
Stearate	3.4e-2	283.2653	1.42	C vs D _ HILIC-	1.24	DOWN

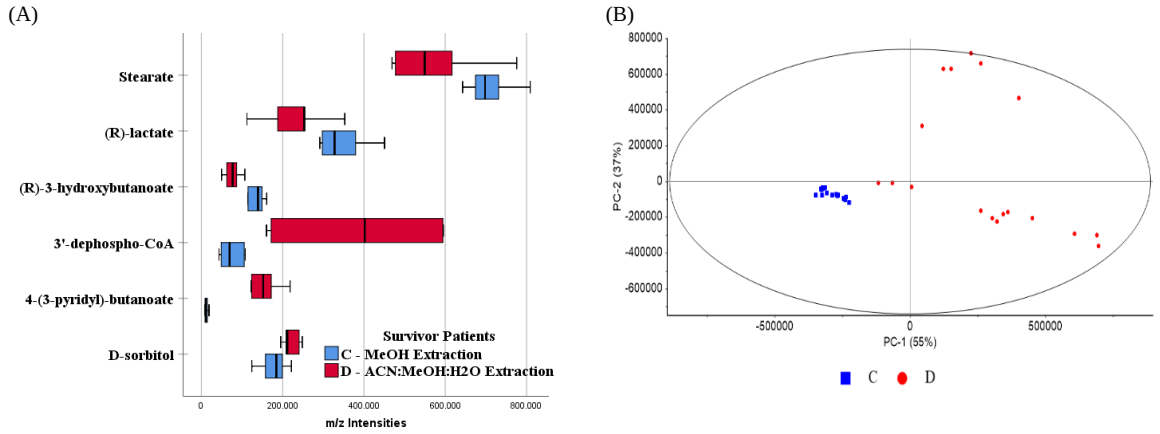


Figure 4.2.1.2 - Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.2. Program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA resultant from allocating the m/z intensity of each metabolite of each group (C: Blue; D: Red) in The Unscrambler® X, is possible to see a fine discernment between group C and group D, being in this case, D better clustered compared to D.

A clear separation between groups is achieved. D-sorbitol, 4-(3-pyridyl)-butanoate and 3'-dephospho-CoA had higher intensities overall in samples submitted to extraction with ACN:MeOH:H_2O. Although the clustering in PCA in figure 4.2.1.2 seems to be better for samples from group C, there is a plausible explanation, since there is no acquisition data for the samples on group D for HILIC column, due to loss of aliquots during the process of injection on the chromatographic columns, making detection of composts like (R)-3-hydroxybutanoate, (R)-lactate, and stearate not possible for 12 samples from group D (ACN:MeOH:H_2O extraction protocol), and 9 samples from group C (methanol extraction).

Eleven distinct metabolites were identified by the two distinct protocols, with molecules from the lipid family (n=3) being heavily up-regulated on samples extracted only with methanol, and molecules composed of nitrogen bonds (n=3) being primarily up-regulated on samples extracted with the ACN:MeOH:H2O protocol, consistent with the results obtained in chapter 4.1.1.

As samples came from patients who survived and patients who died, we tried to determine if there were differences between these outcomes. And so, for the same extraction method, we went for comparing the outcome using the same extraction method, comparing group A vs C (methanol) and group B vs D (ACN:MeOH:H_2O).

Regarding the methanol extraction method, results from the comparison between group A (deceased) and Group C (survivors) are displayed on table 4.2.1.3 and Figure 4.2.1.3.

Table 4.2.1.3 - Metabolites Results from subgroup A vs C. (Abbreviations: R.T: retention time; F.C: fold-change).

Metabolites	*p*-value	*m/z*	R.T. (min)	Pairwise job	F.C	Dysregulation
α-D-glucose	3.1e-7	89.0247	2.09	A vs C _ HILIC-	1.94	DOWN
α-L-iduronate	2.5e-3	193.0358	1.78	A vs C_ C18-	1.78	DOWN
D-sorbitol	2.1e-3	181.0717	1.70	A vs C_ C18-	3.26	DOWN
Biliverdin-IX-α	2.0e-7	583.2571	6.35	A vs C_C18+	2.10	DOWN
Palmitate	6.0e-4	257.2486	13.77	A vs C_C18+	2.23	DOWN
α-linolenate	1.3e-3	279.2328	7.84	A vs C_C18+	1.78	DOWN
Stearate	1.4e-3	285.2801	3.82	A vs C_C18+	1.66	DOWN
4-hydroxy-2-nonenal-glutathione conjugate	1.6e-3	465.2114	11.40	A vs C_C18+	12.12	DOWN
6-pyruvoyl tetrahydropterin	2.1e-3	238.0924	1.60	A vs C_C18+	2.24	DOWN
(4*Z*,7*Z*,10*Z*,13*Z*,16*Z*,19*Z*)-docosahexaenoate	2.1e-3	329.2443	11.09	A vs C_C18+	1.58	UP
4-(3-pyridyl)-butanoate	7.5e-3	166.0849	2.56	A vs C_C18+	2.17	DOWN
N-acetyl-β-neuraminate	8.1e-3	310.1146	1.91	A vs C_C18+	1.45	DOWN
2-trans-hexadecenal	9.3e-3	239.2375	15.52	A vs C_C18+	1.58	DOWN

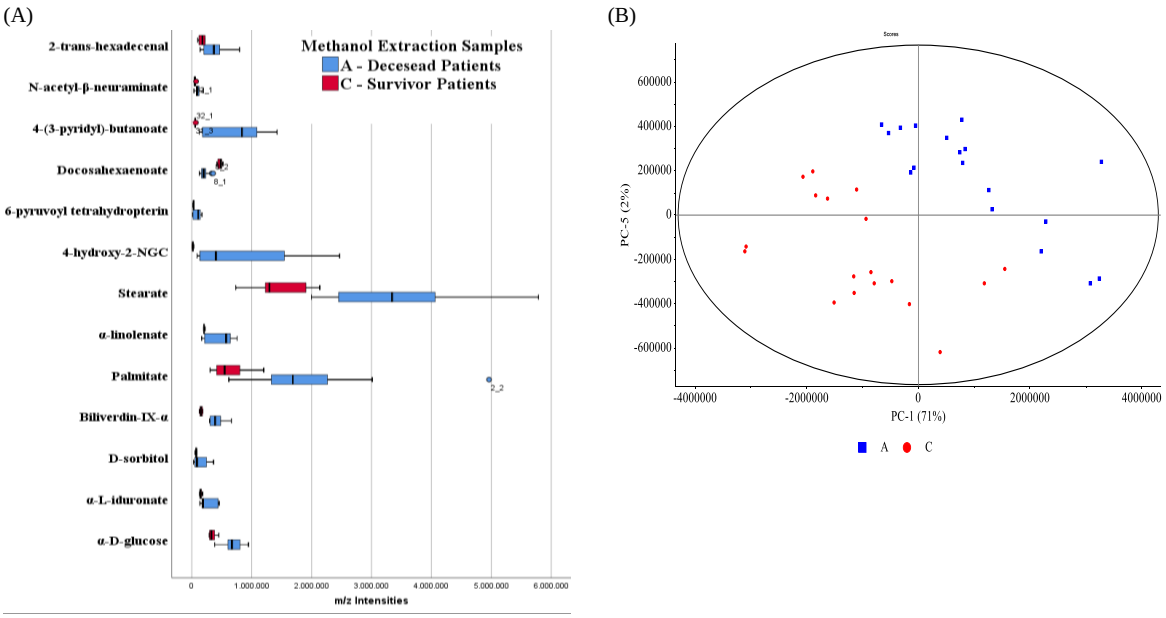


Figure 4.2.1.3 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.3. Each metabolite is represented in both group A (left) and C (right). Y axis is expressed to the power of 0.3. On the X axis each metabolite is represented from left to right by the order indicated on the plot, program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (Abbreviation: 4-hydroxy-2-NGC: 4-hydroxy-2-nonenal-glutathione conjugate). (B) PCA (PC 1 against PC 5) resultant from allocating the m/z intensity of each metabolite of each group (A: Blue; C: Red) in The Unscrambler® X. Although clustering could not be well achieved for both deceased and survivor patient samples, PC1 against PC5 could separate both groups.

As observed on the boxplot from figure 4.2.1.3 (A) *m/z* intensities retrieved from deceased patients' samples (group A) are more dispersed compared to samples from the survivor group (group C). Compared to the PCA (figure 4.2.1.3 (B)) the sample dispersion is relatively the same, achieving only the separation from survivor samples from deceased ones with use of PC1 (x-axis) with PC5 (y-axis).

When using the ACN:MeOH:H_2O extraction samples from deceased patients had, an overall, higher *m/z* intensities for the metabolites found as seen on table 4.2.1.4.

Table 4.2.1.4 - Predictive Metabolites Results from subgroup B vs D. (Abbreviations: R.T: retention time; F.C.: fold-change).

Metabolites	*p*-value	*m/z*	R.T. (min.)	Pairwise job	F.C.	Dysregulation
Hypoxanthine	6.2e-8	135.0302	1.83	B vs D _ C18-	2,04	DOWN
Biliverdin-IX-α	2.8e-7	583.2568	6.40	B vs D _ C18+	2,37	DOWN
D-sorbitol	1.8e-4	183.0854	1.71	B vs D _ C18+	2,65	DOWN
N-formylkynurenine	4.6e-4	238.0926	1.60	B vs D _ C18+	2,85	DOWN
3,4-dihydroxy-5-*all-trans*-decaprenylbenzoate	1.2e-3	835.6668	11.25	B vs D _ C18+	3,12	UP
Sulfate	4.3e-3	96.9604	3.98	B vs D _ C18-	1,59	DOWN
α-L-iduronate	4.8e-3	193.0358	1.78	B vs D _ C18-	2,18	DOWN
4-(3-pyridyl)-butanoate	6.5e-3	166.0848	2.33	B vs D _ C18+	1,90	DOWN
Palmitate	2.0e-2	257.2485	1.04	B vs D _ C18+	1,55	DOWN
N-acetyl-β-neuraminate	3.5e-2	310.1145	1.98	B vs D _ C18+	1,46	DOWN
4α-OCH_3-4β-CH-5α-zymosterol	4.6e-2	429.3743	10.66	B vs D _ C18+	2,04	UP

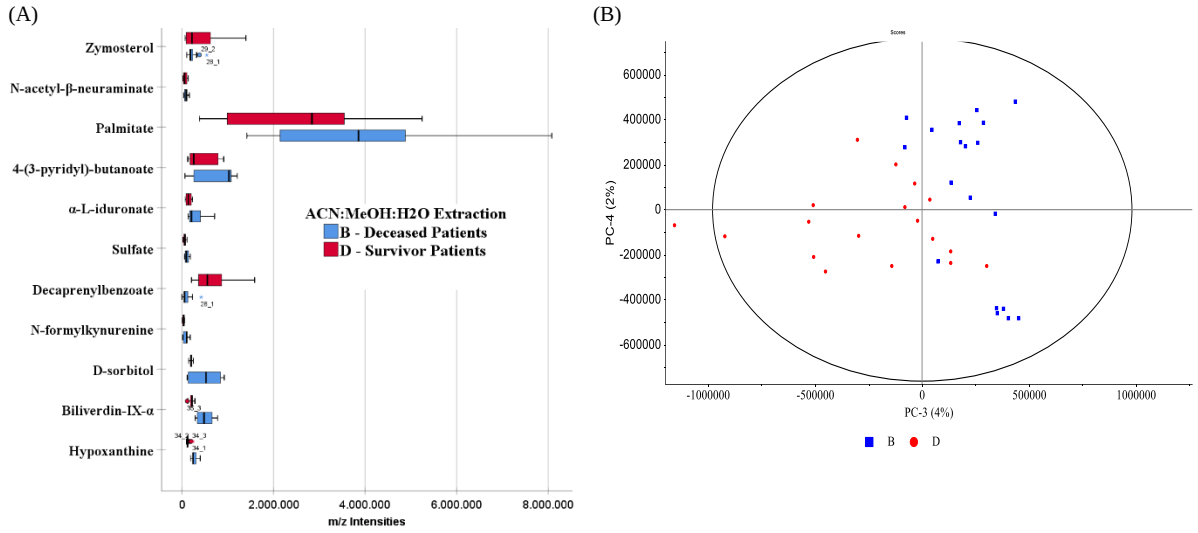


Figure 4.2.1.4 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.1.4. Program used IBM SPSS Statistics. (Abbreviations: Zymosterol: 4α-OCH3-4β-CH-5α-zymosterol; Decaprenylbenzoate: 3,4-dihydroxy-5-all-trans-decaprenylbenzoate). (B) PCA (PC3 against PC4) resultant from allocating the m/z intensity of each metabolite of each group (B: Blue; D: Red) in The Unscrambler®.

ACN:MeOH:H_2O extraction had a worse group separation than Methanol extraction, when observing both PCA's from figure 4.2.1.3(B) and figure 4.2.1.4(B).

4.2.2. Second Assay Results – Metabolome Analysis for Metabolite Identification

From the second set of patients, 24 total patients were selected, 8 who had COVID-19 and were discharged (group A), 8 patients that were submitted to IMV (invasive mechanical ventilation) and got discharged, and 8 patients that were submitted to IMV and deceased on ICU. Samples were treated with the ACN:MeOH:H_2O extraction protocol.

Comparisons were done as followed: Group A versus B comparing, the variable "ventilation" between patients who survived; Group A versus C, comparing outcome between discharged patients and deceased ones; and group B vs C comparing the outcome between patients who had been through ventilation and survived and patients that went through ventilation and died.

Group A vs B comparison retrieved 64 significantly altered features. To reduce this number, features were filtered to eliminate repeated metabolites in different acquisition modes. The minimum *p*-value acceptance was set to ≤7.2e-4, resulting in 29 final unique metabolites. A second filter to the fold-change (fold-change ≥ 3 for both up- and down-regulation) was applied, achieving a final count of 8 metabolites displayed on table 4.2.2.1. figure 4.2.2.1 shows these metabolites on boxplots as well as the PCA based on the metabolites' intensities for the given *m/z*.

Table 4.2.2.1 - Metabolites Results from the comparison between samples from COVID-19 patients who didn't need ventilation and who went through ventilation (group A versus B, respectively). (Abbreviations: R.T: retention time; F.C.: fold-change).

Metabolites	*p*-value	*m/z*	R.T. (min)	Pairwise job	F.C.	Dysregulation
L-ascorbate	1.8e-6	175,03	7,61	A vs B _ HILIC-	4,03	UP
Bilirubin	2.0e-6	585,27	4,50	A vs B _ C18+	3,04	DOWN
Thiamine	6.2e-6	264,11	1,88	A vs B _ HILIC-	16,22	UP
N-acetyl-β-neuraminate	1.3e-5	310,11	1,68	A vs B _HILIC+	9,87	UP
5-methoxytryptophol	3.6e-5	190,09	1,86	A vs B _ HILIC-	4,70	UP
3-keto-L-gulonate	4.8e-5	193,04	9,32	A vs B _ HILIC-	4,28	UP
Calcitriol	7.2e-5	415,32	4,50	A vs B _ C18-	3,35	UP
L-gulonate	2.0e-4	195,05	9,85	A vs B _ HILIC-	5,16	UP

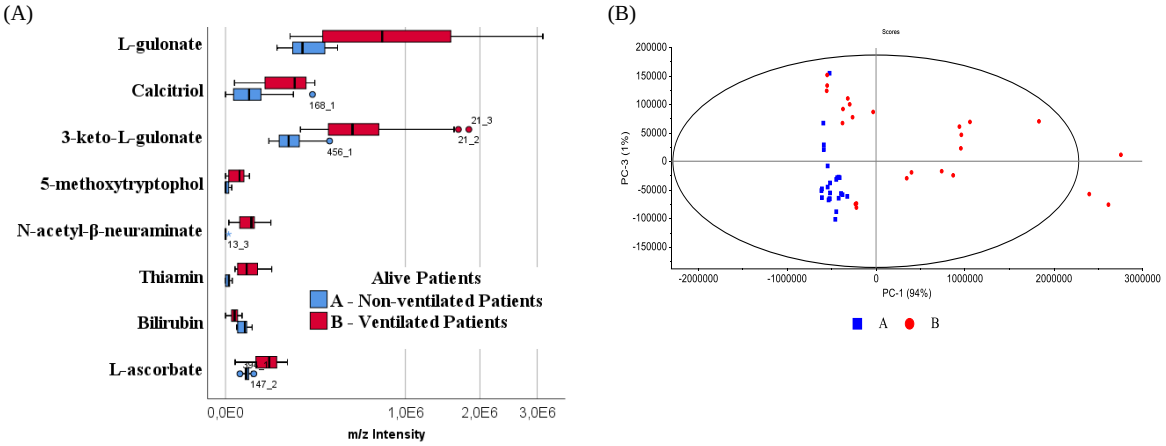
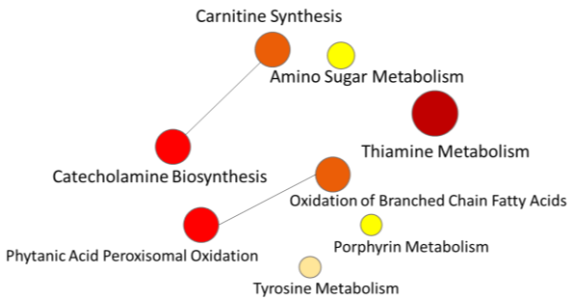


Figure 4.2.2.1 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.1. Y axis is expressed to the power of 0.5. Program used IBM SPSS Statistics. Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC4) resultant from allocating the m/z intensity of each metabolite of each group (A: Alive Non-ventilated patients; B: Alive Ventilated Patients) in The Unscrambler®.

When observing figure 4.2.1.1(B), group separation, with the variable ventilation on check, was well succeeded, patients who didn't go through ventilation clustered better and had lower intensities for the metabolites presented (figure 4.2.2.1(A)). Only one replica from group A (non-ventilated) dispersed from the rest. Ventilated patients (group B) did not cluster onto PCA and had higher m/z intensities for the metabolites identified. These results suggest that patients that required ventilation displayed higher serum levels for the metabolites identified, except for bilirubin.

The HMDB (Human Metabolome Database) ID for the metabolites on table 4.2.2.1 were retrieved and applied on MetaboAnalyst, for Metabolite Set Enrichment Analysis in order to identify the putative metabolic pathways that the metabolites take part [129,130]. The results are shown in the next figure.



Metabolite Set	*p*-value
Thiamine Metabolism	0.0684
Catecholamine Biosynthesis	0.146
Carnitine Synthesis	0.16
Oxidation of Branched Chain Fatty Acids	0.187
Phytanic Acid Peroxisomal Oxidation	0.187
Amino Sugar Metabolism	0.231
Porphyrin Metabolism	0.274
Tyrosine Metabolism	0.443

Figure 4.2.2.2 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between non-ventilated and ventilated alive patients.

Oxidative damage and inflammation are negative events that mutually reinforce each other contributing to inflammation and pro-clotting phenomena. L-ascorbate plays role as a potent antioxidant vitamin in the balance regulation of reactive oxidative species produced during host invasion by infectious targets such as SARS-CoV-2, as well as important co-factor for pathways regulation like the ones mentioned on figure 4.2.2.2. During the acute phases of infection and by consequence inflammation, with no exception for COVID-19, ascorbate levels present in leukocytes and plasma is greatly reduced due to demand, being this molecule a possible predictive of the severity state of the organism. Vollbracht *et al* discusses more in deep the role of the ascorbate in the organism as a response to COVID-19 [131]. Higher levels of this molecule from the ventilated group could mean a positive response to the ventilation and overall therapeutics in reducing the infection and inflammation events.

Bilirubin levels for samples from ventilated patients showed 3 fold-changes higher than the non-ventilated group, this could reflect a potential loss of hepatic function as a detoxification organ, as bilirubin normally binds to glucoronate to be excreted [132]. Hepatic dysfunction may be a predictor of illness severity, which is why patients with greater levels of this molecule might have been in the IMV group, owing to a higher severity condition.

Thiamine (known as vitamin B_1) has a role in immune system function modulation, and in preventing pulmonary oedema working as carbonic anhydrase isoenzyme inhibitor [133]. Ventilated patients had higher levels of this molecule compared to non-ventilated patients, suggesting that thiamine could be involved in the treatment of the IMV submitted patients, since human organism is unable to synthesise this nutrient. Further access to clinical therapeutic is needed to corroborate these findings. Calcitriol, vitamin D activated form founded in the body, also plays a similar role in the protection and balance of the immune system to fight pathogens and prevent autoimmune diseases [134].

N-acetyl-β-neuraminate occurs in glycoproteins/lipids and polysaccharides and is excreted in urine and present in serum. High levels of these molecules in the biofluids mentioned could be indicative of renal failure [135]. Ventilated patients exhibit higher acetyl-neuraminate levels, reflecting a high severity when compared to non-ventilated patients.

5-methoxytryptophol, is an hormone synthetized from the pineal gland with protective effects in terms of antioxidant and immunomodulatory mechanisms [136]. Ventilated patients had the higher levels for this molecule, probably representing the good outcomes as alive discharged patients.

Gulonate may be a by-product of the excretion and metabolization of glucoronate, and the last takes part in the process of glucuronidation that consists in linkage of the glucuronic acid to the substrate (normally therapeutic drugs) which makes the substrate molecules more water-soluble and easier to excrete. This process occurs at the hepatic level.

For comparing groups, non-ventilated alive patients against ventilated deceased patients (group A versus C) the same methodology, as described above, was employed. A total of 52 metabolites (p-value ≤ 8,7E-05) were found, and further filtration to reduce the metabolite number was applied. A p-value ≤ 7.2e-4 cut-off filter was set, leaving

25 metabolites still available. Metabolites with a fold-change lower than 3 for both up- and down-regulation were excluded, leaving 9 metabolites showed on table 4.2.2.2.

Table 4.2.2.2 - Metabolites Results from group A vs C (nonventilated alive patients against ventilated deceased patients). (Abbreviations: R.T: retention time; F.C.: fold-change; FMN: Flavin Mononucleotide).

Metabolites	*p*-value	*m*/z	R.T. (min)	Pairwise job	F.C.	Dysregulation
Octanoate	1,6E-08	145.12	7.38	A vs C _C18+	4,23	UP
Thiamine	8,5E-07	266.12	6.30	A vs C _C18+	7,86	UP
Sucrose	3,7E-06	341.11	1.23	A vs C _C18-	3,84	UP
α-L-iduronate	3,9E-06	193.04	9.31	A vs C _HILIC-	5,22	UP
4-acetamidobutanoate	6,6E-06	144.07	5.71	A vs C _HILIC-	3,52	UP
6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol	2,0E-05	835.69	3.99	A vs C _C18+	3,61	UP
FMN	3,0E-05	455.10	1.20	A vs C _C18-	3,20	UP
3'-dephospho-CoA	5,5E-05	688.16	4.00	A vs C _C18+	3,83	UP
α,α-trehalose	5,5E-05	341.11	9.07	A vs C _HILIC-	5,11	UP

Intensities retrieved from the chosen metabolites were once again grouped up to generate the boxplot and the PCA scatter plot (figure 4.2.2.3).

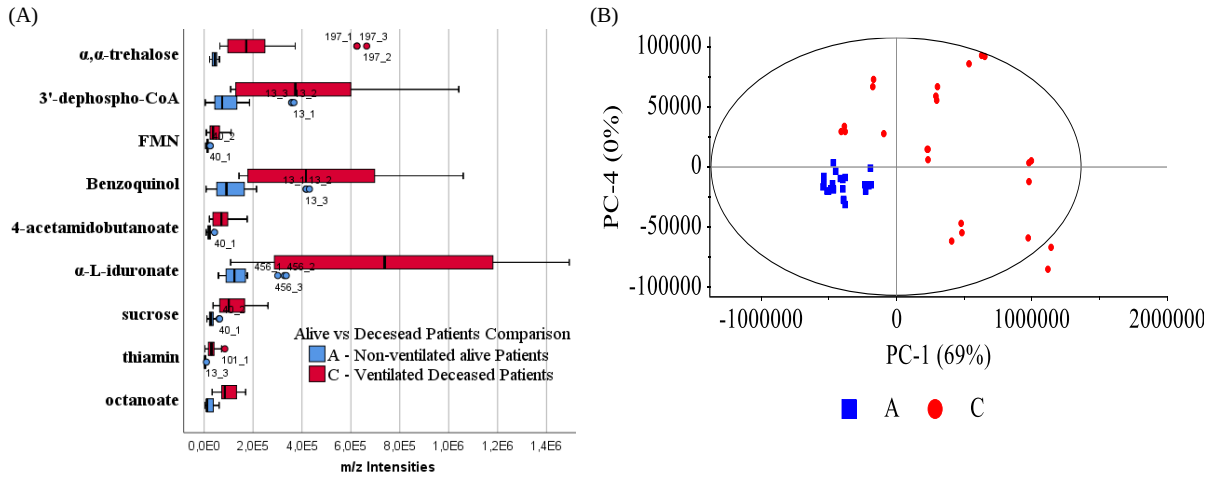
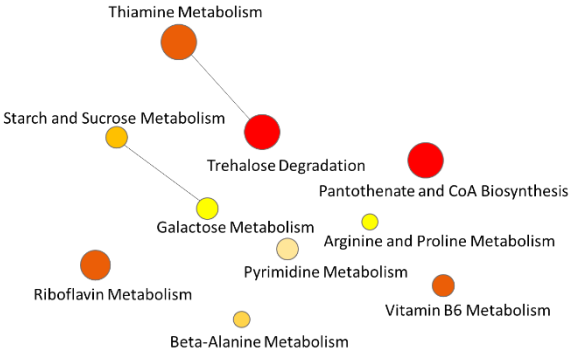


Figure 4.2.2.3 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.2. Program used IBM SPSS Statistics. (Abbreviations: Benzoquinol: 6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol; FMN: Flavin Mononucleotide). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC4) from resultant from allocating the m/z intensity of each metabolite of each group (A: Alive Non-ventilated patients; C: Ventilated deceased patients) in The Unscrambler®.

As observed on table and figure 4.2.2.3 all metabolites are up-regulated for the samples from deceased ventilated patients. PCA from figure 4.2.2.3(B), shows that samples from alive patients are better clustered compared to samples from ventilated deceased patients, possibly related to the fact that *m/z* intensities for the named group are more dispersed as seen on boxplot from figure 4.2.2.3.

The set of metabolites was submitted to enrichment analysis, however HMDB ID's for octanoate and 6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol were not possible to retrieve, and the analysis was followed without these molecules (figure 4.2.2.4).



Metabolite Set	*p*-value
Pantothenate and CoA Biosynthesis	0.00791
Thiamine Metabolism	0.0601
Trehalose Degradation	0.073
Vitamin B6 Metabolism	0.129
Riboflavin Metabolism	0.129
Starch and Sucrose Metabolism	0.194
Beta-Alanine Metabolism	0.211
Galactose Metabolism	0.233
Arginine and Proline Metabolism	0.311
Pyrimidine Metabolism	0.341

Figure 4.2.2.4 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between non-ventilated alive and ventilated deceased patients samples.

The only significant (p-value <0.01) metabolism, uncovered by enrichment analysis was pantothenate (vitamin B5) and CoA biosynthesis modulated by desphospho-CoA and flavin mononucleotide (FMN). This metabolism is mediated by exogenous intake of a micronutrient unable to be synthesized by the organism, vitamin B5, which is a precursor to coenzyme A, essential coenzyme in a variety of biochemical reactions that sustain life, such as TCA cycle, synthesis of fatty acids, steroid hormones, vitamins, acetylcholine (neurotransmitter), and fatty acids beta-oxidation [137]. Deceased patient samples, exhibit higher energy levels for desphospho-CoA molecules compared to alive patient samples, possibly due to therapeutic intake of vitamins or possibly to dysregulation of the normal metabolism caused by SARS-CoV-2 metabolic sequestration.

Thiamine metabolism was identified once again due to the presence of thiamine. Trehalose degradation identified in the presence of α,α-trehalose, from flavin mononucleotide arose vitamin B6, B2, beta-alanine, arginine, proline and pyrimidine metabolism, sucrose was considered part of the starch and sucrose, and galactose metabolism, with substantially low p-values for metabolic processes considered.

Even though these metabolic processes were not significant the metabolites are and take part on metabolism than those identified and should be considered as biomarkers candidates to distinguish outcome between COVID-19 patients.

Octanoate , is a medium-chain free fatty acid that is mainly beta-oxidized on the liver on its free form and oxidized on its carnitine bound form on the skeletal muscle and myocardium. Since is related to the mitochondria energetic catabolism, this molecule could be indicator of any energetic dysregulation or potential organ failure [138]. Deceased patients sera had higher concentrations of free octanoate compared to survived patients.

Vitamin B_1 (thiamine), was once again found dysregulated, with higher levels for deceased patients. Possible cause for this phenomenon, could be in the therapeutical intake of this vitamin to ameliorate the patients' condition. Further study of patients clinical data is advised.

Sucrose, as well trehalose, are disaccharides compost of fructose and glucose, and two molecules of glucose respectively. Normally these sugars are break down into monosaccharides in the intestinal wall before reaching the bloodstream. Higher levels of this molecules on the serum, could demonstrate an impairment function of the enterocytes in producing the enzymes capable of break the sucrose [139]. In addition to this, if small intestine dysfunction caused by SARS-CoV-2, could make higher levels of sugar in blood, could be an addition factor for immune response exacerbation, since the glucose metabolism is said to be interconnected with the cytokine storm response [140,141]. The higher levels of sucrose molecules in the deceased patient samples, could be leading of a poor prognosis.

Iduronate is a constituent of glycosaminoglycans, which could derived from the heparin therapy, as the patients undergo heavy anticoagulation therapy [142,143].

Acetamidobutanoate is a GABA (gamma aminobutyric acid) derivate, a product of the urea cycle and amino acid metabolism such as proline and arginine, and higher levels of this molecule could mean a high metabolic activity by patients in the deceased group.

6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol is a hydroquinone that participates in the metabolism of vitamins such as riboflavin, which in turn is a precursor of metabolites like flavin mononucleotide, essential co-factor for processes like beta-oxidation, TCA cycle, and other catabolic processes. Having high levels of these two mentioned molecules on the deceased patient sera, it's a strong prove that metabolic processes are at high expression.

Desphospho-CoA is precursor to the coenzyme A, and a metabolite part of the pantothenate (vitamin B_5) metabolism. High levels of expression could be once again indicator of a high metabolic activity remarked on patients from the deceased group.

A third comparison, between survivors and deceased patients, that been both trough ventilation (group B vs C), was performed, and 64 significantly altered metabolites were identified. Repeated metabolites were once more filtered, leaving 39 unique metabolites. *P*-value ≤7.2e-4 filter was applied, and 6 metabolites were left. When applied the fold-change ≥3 filter, no metabolites would be left, so this step was leaped. Table 4.2.2.3 shows the metabolites selected.

Table 4.2.2.3 - Metabolites Results from subgroup B vs C. (Abbreviations: R.T: retention time; F.C.: fold-change).

Metabolites	*p*-value	*m*/z	R.T. (min)	Pairwise job	F.C.	Dysregulation
Phenylacetate	3,10E-06	135,045	1,82	B vs C _HILIC-	1,35	DOWN
4-(3-pyridyl)-butanoate	9,40E-06	166,09	1,58	B vs C _C18+	1,29	DOWN
5,6-dihydrouracil	2,90E-04	113,04	9,09	B vs C _HILIC-	1,84	UP
Stearate	3,10E-04	283,26	1,69	B vs C _HILIC-	1,31	DOWN
Linoleate	5,80E-04	281,25	1,73	B vs C _HILIC+	1,74	DOWN
Octanoate	5,90E-04	145,12	7,37	B vs C _C18+	1,62	UP

Fold-change from the most significant metabolites were low compared to previous results. Deceased patients' samples had higher intensity dispersion for the significant altered metabolites set (figure 4.2.2.5(A)). PCA scatter plot did not achieve a cluster between group B and C, but there is a tendency for samples from alive ventilated patients to be allocated on the 1st and 2nd quadrants of the PCA, while deceased ventilated patients tend to be allocated on the 3rd and 4th quadrant (figure 4.2.2.5(B)).

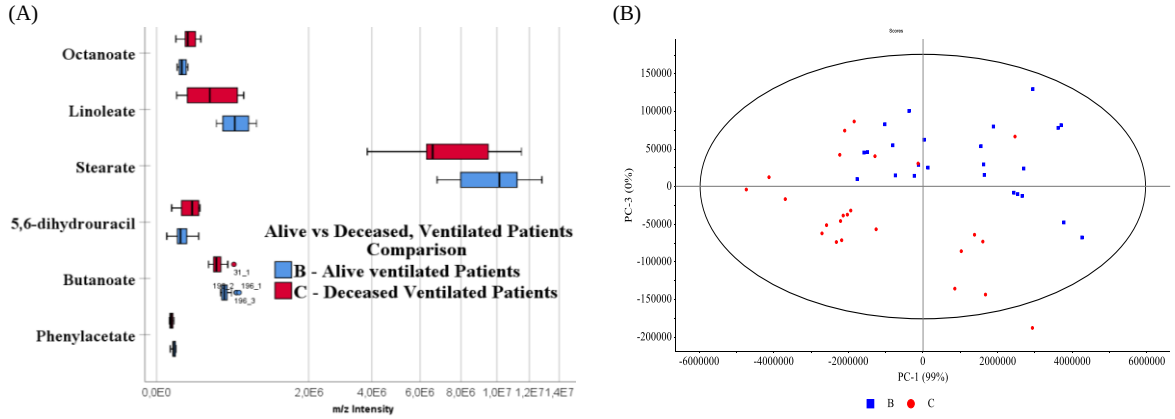
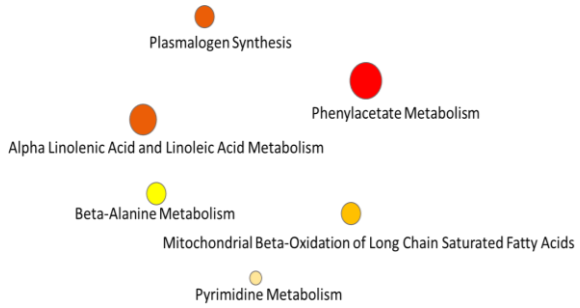


Figure 4.2.2.5 – (A) Box-Plot generated by the m/z intensities for each metabolite expressed in table 4.2.2.3. Intensity for the given metabolite (y-axis) is expressed to the power of 0.5. Program used IBM SPSS Statistics. (Abbreviations: Butanoate: 4-(3-pyridyl)-butanoate). Extracted ion chromatograms (EIC) for all metabolites here represented will be placed in the Appendix. (B) PCA (PC1 against PC3) resultant from allocating the m/z intensity of each metabolite of each group (B: Alive Ventilated patients' samples; C: Deceased Ventilated patients' samples) in The Unscrambler®.

Enrichment analysis, with exception of octanoate was executed, results on figure 4.2.2.6.



Metabolite Set	*p*-value
Phenylacetate Metabolism	0.0433
Alpha Linolenic Acid and Linoleic Acid Metabolism	0.0896
Plasmalogen Synthesis	0.121
Mitochondrial Beta-Oxidation of Long Chain Saturated Fatty Acids	0.13
Beta-Alanine Metabolism	0.156
Pyrimidine Metabolism	0.257

Figure 4.2.2.6 - Metabolite Set Enrichment Analysis from the metabolites retrieved from group comparison between ventilated alive and deceased patients samples.

The phenylacetate metabolism, identified with the contribution of phenylacetate, was the only statistically valid metabolism. Because phenylacetate metabolism promotes nitrogen excretion in individuals with urea cycle abnormalities, it is significant for therapeutic purposes [144]. Also, concentration shift of this molecules could be possible due to COVID-19 cytopathic renal damage and ECMO therapy in the patients, that in addition to blood oxygenation it functions as a haemodialysis system, altering the blood levels of this molecule. Further information about the patients regarding the procedures executed are needed to evaluate these results.

Linoleate was considerate to be involved in the metabolism of Alpha Linolenic Acid and Linoleic Acid Metabolism. The remaining metabolisms with substantially higher p-values were, Plasmalogen Synthesis, Mitochondrial Beta-Oxidation of Long Chain Saturated Fatty Acids, with contribution from the stearate molecule identified with higher energy intensities for the ventilated alive patients, possible indicative of higher energetic catabolism for the ventilated group. Beta-alanine metabolism and pyrimidine metabolism were considered due to existence of 5,6-dihydrouracil which was in a higher intensity state on ventilated deceased patients.

4-(3-pyridyl)-butanoate, is a pyridine derivate with few information available, being only associated with nicotine degradation and involved in smoking and lung cancer studies [145]. Despite the levels of this metabolite being higher for the ventilated alive group, only two patients in this study were reported to be smokers, being inserted on the deceased patients group.

Dihydrouracil (5,6-Dihydrouracil) is an uracil metabolite that has been used to detect dihydropyrimidine dehydrogenase (DPD)-deficient individuals. DPD activity deficiencies relate to decreased 5,6-Dihydrouracil catabolism, which can result in severe toxicity in a variety of clinical situations. Because DPD deficiency inhibits the metabolic breakdown of 5,6-Dihydrouracil, the build-up of this uracil in the plasma of individuals with clinical conditions can be recognised [146]. Higher levels of this molecule on deceased patients could be indicative of a metabolic disorder related to liver function or even peripheral mononuclear cells function impairment, since that dihydrouracil conversion occurs only in these two cell types that possess DPD enzyme.

Octanoate was found up-regulated on the deceased patients' group, as a free fatty acid, it is only oxidized at the liver mitochondria, and higher blood levels of this molecule might be indicative of organ dysfunction, namely at liver level [138].

Machine learning algorithms applied to mass spectrometric data collected from the 3 groups comparison achieved excellent results, table below shows performance parameters and algorithms used.

Table 4.2.2.4 – Machine learning results from samples retrieved from mass spectrometer data from the second assay patients' samples. Models computed to determine in each data set the patient submission to ventilation (group A vs B); outcome between non-ventilated and ventilated deceased patients (group A vs C); outcome between ventilated patients (Group B vs C). Sampling type: Leave One Out (LOO); Evaluation results for target: average over class. Models were run on default setup offered by Orange 3.32 software. (Abbreviations: SVM: Support Vector Machine; kNN: k-nearest neighbours algorithm; SGD: Stochastic Gradient Descent; AUC: Area Under the Curve; CA: Classification Accuracy; F1: F1-score is the harmonic mean of precision and recall).

	A vs B					A vs C					B vs C				
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**AUC**	**CA**	**F1**	**Precision**	**Recall**
SVM	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.96	0.96	0.96	0.96	0.96
Neural Network	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.96	0.96	0.96	0.96
AdaBoost	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.90	0.90	0.90	0.90	0.90
kNN	0.983	0.875	0.874	0.886	0.875	0.99	0.90	0.90	0.91	0.90	0.78	0.75	0.75	0.75	0.75
SGD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.96	0.96	0.96

Overall predictions from the mass spectrometry data could distinguish samples from each group. The best algorithm to achieve the results desired was the neural network. Only comparation between the ventilated discharged patients and ventilated deceased patients (group B vs group C) had two misclassified samples, corresponding to one replica from patient 21 and one replica from the patient 31. Confusion matrix table is shown below (table 4.2.2.5)

Table 4.2.2.5 – Confusion matrix retrieved from neural network. Ventilated discharged patients versus ventilated deceased patients (group B vs C). Percentages relative to proportion of predicted samples.

		Predicted		
	Outcome	Survival	Death	Σ
Actual	Survival	95.8%	4.2%	24
	Death	4.2%	95.8%	24
	Σ	24	24	48

The set of metabolites chosen for each comparison were enough to distinguish between each patient sample condition and/or outcome, revealing that these molecules can work as future biomarkers for prediction of future patients' condition and outcome on ICU admission.

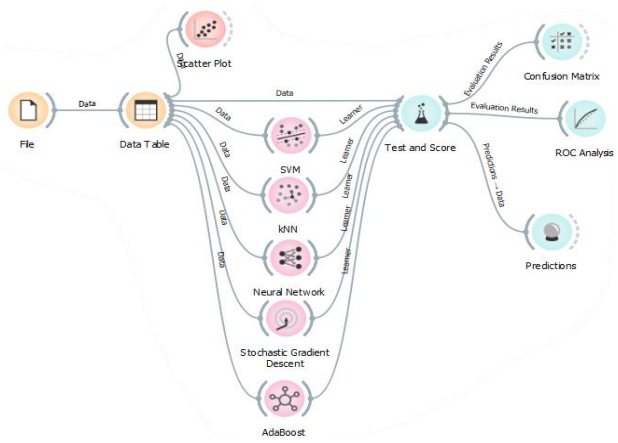


Figure 4.2.2.7 – workflow for learning and prediction analysis for chapter 4.2.2.

Chapter 5 – Conclusion

It can be concluded that in the 1st assay for testing the two extraction protocols influence on the metabolites analysed the methanol extraction had a lipidic signature patent on FTIR analysis by the wavenumbers 2924 cm^{-1} and 2853 cm^{-1}. These results are concordant with the LC/MS analysis, where it was found that lipid molecules (stearate, 4α-OCH3-4β-CH-5α-zymosterol, 2-trans-hexadecenal, and α-tocopherol) were up-regulated, with a fold-change between 1.24 and 9, higher for the methanol protocol. On the other hand, ACN:MeOH:H_2O had a FTIR signature more relatable to peptide bonds, with influence of the wavenumbers 1655 cm^{-1} and 1556 cm^{-1}. Results from mass spectrometry were concordant as well, the ACN:MeOH:H_2O protocol revealed higher values for purine and pyridine like molecules (3'-dephospho-CoA, 4-(3-pyridyl)-butanoate) with fold-changes between 8.86 and 16.31.

Regardless of the extraction protocol, prediction for outcome (dead/alive) reached an AUC of 0.98 with the Stochastic Gradient Descent, with FTIR spectral data.

For the 2nd essay, the sample pool size was raised to 8 individuals per group, with the addition of one condition, invasive mechanical ventilation, giving raise to three distinct groups, non-ventilated discharged patients, ventilated discharged patients and ventilated deceased patients, all COVID-19 positive. Using the FTIR Spectral data, predictions were generated for each comparison, and AUCs of 0.84 for ventilation between discharged patients, 0.93 between discharged and deceased patients, and 0.79 between discharged ventilated and deceased ventilated patients were obtained. Using LC-MS data for the same patients, AUCs of 1.00 (for ventilation status and outcome between non-ventilated and ventilated deceased patients) and 0.97 were achieved for the three comparisons. Despite the better results for LC-MS data, the effort to achieve these grades are far higher than FTIR analysis, since the time it takes to perform a LC-MS analysis is scaled to hours or even days depending on the sample size, and FTIR analysis takes seconds to minutes for the same sample size. Data from LC-MS must first be converted to a format that is compatible with the software used for MS analysis and then uploaded to that platform for metabolite investigation, and only after that is data suitable for inquiry and conclusion. Even so LC-MS data has higher quality when defining what metabolites are at each sample, as on the second assay it was possible to observe what metabolites were up- and down- regulated for each group, and it was possible to determine the putative metabolic pathways associated to each comparison. The metabolites collected were mostly connected with infection and inflammatory metabolism, such as dysregulation at the energetic catabolic level (TCA cycle and beta-oxidation of free fatty acids). Metabolites were discovered to be connected with the metabolism of potentially therapeutic molecules such as vitamins and elements related to hepatic and renal dysfunction.

The collective metabolites were predictive of outcomes and ventilation status across all comparisons, demonstrating a viable prospective biomarker for COVID-19 outcome prediction. FTIR analysis has a great potential to follow the patients' outcome, with promising prediction values and a much lower sample time analysis.

Chapter 6 – Future Work & Perspectives

In this thesis, we addressed the problem of biomarker discovery for COVID-19 based on conventional and non-conventional metabolomic platforms, mass spectrometry and Fourier-transform infrared spectroscopy. One of the main contributions of this work was to address the impact of the solvent used to extract the desired metabolites, as one extraction method only with methanol yields a higher lipidic content and another extraction method with acetonitrile, methanol, and water yields a molecular profile rich in peptide family metabolites. On the other hand, both techniques had the capacity to identify differences between the established groups, proving that the biomarkers identified were of value, being on the FTIR case the wavenumbers related to the peptide bonds and possibly RNA/DNA bonds, and in the case of the mass spectrometry data the metabolites with significant dysregulation levels.

Globally, the results obtained were promising. FTIR data was capable of distinguish deceased patients samples from discharged patients samples, as observed on chapter 4.1.1, as well as to observe the impact on the solvent or mixture of solvents used for the samples. On chapter 4.1.2 the aim was more ambitious, as a new condition status was added, invasive mechanical ventilation. Even so, the prediction algorithms were able to distinguish between ventilated from non-ventilated patients and outcome (discharged/ deceased).

In order to better understand the sample composition conventional metabolomics was employed, by resolving to LC/MS. The number of instances studied were far less when compared to the 400 cm^{-1} to 4000 cm^{-1}, which facilitated analysis and a better comprehension of the metabolic state of each group. However, mass spectrometry is an extensive and slow platform, compared to FTIR analysis. MS analysis is scaled in hours to days when analysing a small population like in this work (a total of 30 patients were studied), when FTIR takes seconds to minutes to analyse the same number of samples.

Notwithstanding, there is still much that can be done in this work. The main obstacle that this work faced was the sample size. Each group contained from 3 or 8 patients, which poses a serious risk of not representing the population of the Hospital São José. Regardless, the metabolites discovered had statistical value and for the samples chosen it was able to predict the outcome and the ventilation condition for the patients. A future perspective is to extend the population size with the goal of creating a realistic database, in terms of size, with FTIR serum profiles, and metabolomics data, capable of predicting patient prognosis in a fast, simple and systematic way, paving the way for a new tool in the help for treatment and care of patients.

From an experimental standpoint, the contributions of this work lie in the development of a work methodology that can fit the everyday workflow in an analysis laboratory, as demonstrated here by the transition from a wet lab to an exploratory data analysis with the interconnected use of various programmes such as, XCMS, MetaboAnalyst 5.0, The Unscrambler X, IBM SPSS Statistics, and ORANGE Data Mining 3.3.

Regardless of the performance of the algorithms used and the approaches taken, the following ideas could be tested:

1. Design of a unified model including both FTIR and MS data, to combine both platforms and obtain one model with higher accuracy and performance.
2. It would be easier to understand and interpret the data acquired if the clinical factors were taken into consideration and related to the metabolites discovered, this would only be possible by having a group force working on, like in the PREMO Project.

As part of the PREMO project, it is heartening to state that this study can provide some perspectives for continuing the development and validation of novel ways of interpreting data and perhaps using them in a hospital setting.

References

1. Araújo R, Bento LFN, Fonseca TAH, Von Rekowski CP, da Cunha BR, Calado CRC. Infection Biomarkers Based on Metabolomics. Metabolites. 2022 Jan 19;12(2):92. DOI: 10.3390/metabo12020092

2. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology. 2019;17(3):181–92. DOI: 10.1038/s41579-018-0118-9

3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020 Feb 20;382(8):727–33. DOI: 10.1056/NEJMoa2001017

4. Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. Viruses. 2020 Jan 24;12(2):135. DOI: 10.3390/v12020135

5. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology. 2021;19(3):141–54. DOI: 10.1038/s41579-020-00459-7

6. Team E editorial. Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern. Eurosurveillance. 2020 Feb 6 [cited 2021 Aug 31];25(5):200131e. DOI: 10.2807/1560-7917.ES.2020.25.5.200131e

7. World Health Organization. Novel Coronavirus(2019-nCoV) : Situation Report – 22. 2020 [cited 2021 Aug 31].

8. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology. 2020 Apr 2 [cited 2021 Aug 31];5(4):536–44. DOI: 10.1038/s41564-020-0695-z

9. World Health Organization. Coronavirus Disease 2019 Situation Report – 51. 2020 [cited 2021 Aug 31];14(6).

10. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine. 2020 Mar 26;382(13):1199–207. DOI: 10.1056/NEJMoa2001316

11. Masters PS. The Molecular Biology of Coronaviruses. In 2006. p. 193–292. DOI: 10.1016/S0065-3527(06)66005-3

12. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, Yuen KY. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes & Infections. 2020 Jan 1;9(1):221–36. DOI: 10.1080/22221751.2020.1719902

13. Woo PCY, Lau SKP, Lam CSF, Lau CCY, Tsang AKL, Lau JHN, Bai R, Teng JLL, Tsang CCC, Wang M, Zheng BJ, Chan KH, Yuen KY. Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavi. Journal of virology. 2012 Apr;86(7):3995–4008. DOI: 10.1128/JVI.06540-11

14. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020;395(10224):565–74. DOI: 10.1016/S0140-6736(20)30251-8

15. Mittal A, Manjunath K, Ranjan RK, Kaushik S, Kumar S, Verma V. COVID-19 pandemic: Insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2. PLOS Pathogens. 2020 Aug 21;16(8):e1008762. DOI: 10.1371/journal.ppat.1008762

16. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar 12;579(7798):265–9. DOI: 10.1038/s41586-020-2008-3

17. Yao H, Song Y, Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, Zhang J. Molecular Architecture of the SARS-CoV-2 Virus. Cell. 2020;183(3):730-738.e13.

18. Brant AC, Tian W, Majerciak V, Yang W, Zheng ZM. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell and Bioscience. 2021;11(1):1–17. DOI: 10.1186/s13578-021-00643-z

19. Wang MY, Zhao R, Gao LJ, Gao XF, Wang DP, Cao JM. SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. Frontiers in Cellular and Infection Microbiology. 2020 Nov 25;10(November):1–17. DOI: 10.3389/fcimb.2020.587269

20. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nature Communications. 2021;12(1):1–20. DOI: 10.1038/s41467-021-22905-7

21. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature. 2021;589(7840):125–30. DOI: 10.1038/s41586-020-2739-1

22. To KKW, Sridhar S, Chiu KHY, Hung DLL, Li X, Hung IFN, Tam AR, Chung TWH, Chan JFW, Zhang AJX, Cheng VCC, Yuen KY. Lessons learned 1 year after SARS-CoV-2 emergence leading to COVID-19 pandemic. Emerging Microbes & Infections. 2021 Jan 1;10(1):507–35. DOI: 10.1080/22221751.2021.1898291

23. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, Atif SM, Hariprasad G, Hasan GM, Hassan MI. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2020 Oct;1866(10):165878. DOI: 10.1016/j.bbadis.2020.165878

24. Yadav R, Chaudhary JK, Jain N, Chaudhary PK, Khanra S, Dhamija P, Sharma A, Kumar A, Handu S. Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. Cells. 2021 Apr 6;10(4):821. DOI: 10.3390/cells10040821

25. Raj R. Analysis of non-structural proteins, NSPs of SARS-CoV-2 as targets for computational drug designing. Biochemistry and Biophysics Reports. 2021;25:100847. DOI: 10.1016/j.bbrep.2020.100847

26. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020 May 14 [cited 2021 Sep 6];181(4):914-921.e10. DOI: 10.1016/j.cell.2020.04.011

27. Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom KJ, Zambon M, Ellis J, Lewis PA, Hiscox JA, Matthews DA. Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Medicine. 2020 Dec 28;12(1):68. DOI: 10.1186/s13073-020-00763-0

28. Yeung ML, Teng JLL, Jia L, Zhang C, Huang C, Cai JP, et al. Soluble ACE2-mediated cell entry of SARS-CoV-2 via interaction with proteins related to the renin-angiotensin system. Cell. 2021 Apr;184(8):2212-2228.e12. DOI: 10.1016/j.cell.2021.02.053

29. Wang K, Chen W, Zhang Z, Deng Y, Lian JQ, Du P, et al. CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. Signal Transduction and Targeted Therapy. 2020;5(1):1–10. DOI: 10.1038/s41392-020-00426-x

30. Wei C, Wan L, Yan Q, Wang X, Zhang J, Yang X, et al. HDL-scavenger receptor B type 1 facilitates SARS-CoV-2 entry. Nature Metabolism. 2020;2(12):1391–400. DOI: 10.1038/s42255-020-00324-0

31. Andersson U, Ottestad W, Tracey KJ. Extracellular HMGB1: a therapeutic target in severe pulmonary inflammation including COVID-19? Molecular Medicine. 2020 Dec 7 [cited 2021 Sep 7];26(1):42. DOI: 10.1186/s10020-020-00172-4

32. Wei J, Alfajaro MM, DeWeirdt PC, Hanna RE, Lu-Culligan WJ, Cai WL, et al. Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. Cell. 2021 Jan;184(1):76-91.e13. DOI: 10.1016/j.cell.2020.10.028

33. Clausen TM, Sandoval DR, Spliid CB, Pihl J, Perrett HR, Painter CD, et al. SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. Cell. 2020 Nov 12 [cited 2021 Oct 18];183(4):1043-1057.e15. DOI: 10.1016/j.cell.2020.09.033

34. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020 Mar 30 [cited 2021 Oct 20];581(7807):221–4. DOI:

10.1038/s41586-020-2179-y

35. Bestle D, Heindl MR, Limburg H, Van Lam van T, Pilgram O, Moulton H, et al. TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells. Life Science Alliance. 2020 Sep 23;3(9):e202000786. DOI: 10.26508/lsa.202000786

36. Vaarala MH, Porvari KS, Kellokumpu S, Kyllönen AP, Vihko PT. Expression of transmembrane serine protease TMPRSS2 in mouse and human tissues. The Journal of pathology. 2001 Jan [cited 2021 Oct 17];193(1):134–40. DOI: 10.1002/1096-9896(2000)9999:9999<::AID-PATH743>3.0.CO;2-T

37. Chen YW, Lee MS, Lucht A, Chou FP, Huang W, Havighurst TC, Kim K, Wang JK, Antalis TM, Johnson MD, Lin CY. TMPRSS2, a Serine Protease Expressed in the Prostate on the Apical Surface of Luminal Epithelial Cells and Released into Semen in Prostasomes, Is Misregulated in Prostate Cancer Cells. The American Journal of Pathology. 2010 Jun [cited 2021 Oct 17];176(6):2986–96. DOI: 10.2353/ajpath.2010.090665

38. Glowacka I, Bertram S, Muller MA, Allen P, Soilleux E, Pfefferle S, et al. Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response. Journal of Virology. 2011 May 1 [cited 2021 Oct 17];85(9):4122–34. DOI: 10.1128/jvi.02232-10

39. Zang R, Castro MFG, McCune BT, Zeng Q, Rothlauf PW, Sonnek NM, et al. TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. Science Immunology. 2020 May 13 [cited 2021 Oct 17];5(47). DOI: 10.1126/sciimmunol.abc3582

40. Snijder EJ, Limpens RWAL, de Wilde AH, de Jong AWM, Zevenhoven-Dobbe JC, Maier HJ, Faas FFGA, Koster AJ, Bárcena M. A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. PLOS Biology. 2020 Jun 8;18(6):e3000715. DOI: 10.1371/journal.pbio.3000715

41. Klein S, Cortese M, Winter SL, Wachsmuth-Melm M, Neufeldt CJ, Cerikan B, Stanifer ML, Boulant S, Bartenschlager R, Chlanda P. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. Nature Communications. 2020;11(1):1–10. DOI: 10.1038/s41467-020-19619-7

42. de Wilde AH, Snijder EJ, Kikkert M, van Hemert MJ. Host factors in coronavirus replication. InSpringer, ChamSpringer, Cham Springer, ChamSpringer, Cham 2018 [cited 2021 Oct 18]. p. 1–42. DOI: 10.1007/82_2017_25

43. Burmer G, Burmer M, Pabuwal V. SARS-CoV-2 and COVID-19 Pathogenesis: A Review | LSBio. LifeSpan BioSciences, Inc. [cited 2022 May 31].

44. Hossain MU, Bhattacharjee A, Emon MTH, Chowdhury ZM, Ahammad I, Mosaib MG, et al. Novel mutations in NSP-1 and PLPro of SARS-CoV-2 NIB-1 genome mount for effective therapeutics. Journal of Genetic Engineering and Biotechnology. 2021 Dec 2;19(1):52. DOI: 10.1186/s43141-021-00152-z

45. Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ. Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling. Journal of Virology. 2009 Oct;83(19):10314–8. DOI: 10.1128/JVI.00842-09

46. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Research. 2018 Jan;149(January):58–74. DOI: 10.1016/j.antiviral.2017.11.001

47. Chakraborty J, Maity A, Sarkar H. A systematic drug repurposing approach to identify promising inhibitors from FDA-approved drugs against Nsp4 protein of SARS-CoV-2. Journal of Biomolecular Structure and Dynamics. 2021 Nov 30;0(0):1–10. DOI: 10.1080/07391102.2021.2009033

48. Chen J, Li Z, Guo J, Xu S, Zhou J, Chen Q, Tong X, Wang D, Peng G, Fang L, Xiao S. SARS-CoV-2 nsp5 Exhibits Stronger Catalytic Activity and Interferon Antagonism than Its SARS-CoV Ortholog. Journal of Virology. 2022 Apr 27;96(8). DOI: 10.1128/jvi.00037-22

49. Sun X, Liu Y, Huang Z, Xu W, Hu W, Yi L, et al. SARS-CoV-2 non-structural protein 6 triggers NLRP3-dependent pyroptosis by targeting ATP6AP1. Cell Death & Differentiation. 2022 Jan 8;(June). DOI: 10.1038/s41418-021-00916-7

50. Reshamwala SMS, Likhite V, Degani MS, Deb SS, Noronha SB. Mutations in SARS-CoV-2 nsp7 and nsp8 proteins and their predicted impact on replication/transcription complex structure. Journal of Medical Virology. 2021 Jul 14;93(7):4616–9. DOI: 10.1002/jmv.26791

51. Konkolova E, Klima M, Nencka R, Boura E. Structural analysis of the putative SARS-CoV-2 primase complex. Journal of Structural Biology. 2020 Aug;211(2):107548. DOI: 10.1016/j.jsb.2020.107548

52. de O. Araújo J, Pinheiro S, Zamora WJ, Alves CN, Lameira J, Lima AH. Structural, energetic and lipophilic analysis of SARS-CoV-2 non-structural protein 9 (NSP9). Scientific Reports. 2021 Dec 26;11(1):23003. DOI: 10.1038/s41598-021-02366-0

53. Benoni R, Krafcikova P, Baranowski MR, Kowalska J, Boura E, Cahov H. Substrate Specificity of SARS-CoV-2 Nsp10-Nsp16 Methyltransferase. 2021;

54. Jing H, Ke W, Tao R, Li Y, Zhao Y, Cao S, Sun Y, Wang J, Zhang Y, Dong W, Zhao P, Duan E. TRIM59 inhibits porcine reproductive and respiratory syndrome virus (PRRSV)-2 replication in vitro. Research in Veterinary Science. 2019 Dec;127(April):105–12. DOI: 10.1016/j.rvsc.2019.10.004

55. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nature Communications. 2019 Dec 28;10(1):2342. DOI: 10.1038/s41467-019-10280-3

56. Shu T, Huang M, Wu D, Ren Y, Zhang X, Han Y, Mu J, Wang R, Qiu Y, Zhang DY, Zhou X. SARS-Coronavirus-2 Nsp13 Possesses NTPase and RNA Helicase Activities That Can Be Inhibited by Bismuth Salts. Virologica Sinica. 2020 Jun 4;35(3):321–9. DOI: 10.1007/s12250-020-00242-1

57. Newman JA, Douangamath A, Yadzani S, Yosaatmadja Y, Aimon A, Brandão-Neto J, Dunnett L, Gorrie-stone T, Skyner R, Fearon D, Schapira M, von Delft F, Gileadi O. Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. Nature Communications. 2021 Dec 11;12(1):4848. DOI: 10.1038/s41467-021-25166-6

58. Martin WR, Cheng F. Repurposing of FDA-Approved Toremifene to Treat COVID-19 by Blocking the Spike Glycoprotein and NSP14 of SARS-CoV-2. Journal of Proteome Research. 2020 Nov 6;19(11):4670–7. DOI: 10.1021/acs.jproteome.0c00397

59. Frazier MN, Dillard LB, Krahn JM, Perera L, Williams JG, Wilson IM, Stewart ZD, Pillon MC, Deterding LJ, Borgnia MJ, Stanley RE. Characterization of SARS2 Nsp15 nuclease activity reveals it's mad about U. Nucleic Acids Research. 2021 Sep 27;49(17):10136–49. DOI: 10.1093/nar/gkab719

60. Vithani N, Ward MD, Zimmerman MI, Novak B, Borowsky JH, Singh S, Bowman GR. SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential. Biophysical Journal. 2021 Jul;120(14):2880–9. DOI: 10.1016/j.bpj.2021.03.024

61. Baggen J, Vanstreels E, Jansen S, Daelemans D. Cellular host factors for SARS-CoV-2 infection. Nature Microbiology. 2021 Oct 1;6(10):1219–32. DOI: 10.1038/s41564-021-00958-0

62. Wolff G, Limpens RWAL, Zevenhoven-Dobbe JC, Laugks U, Zheng S, de Jong AWM, Koning RI, Agard DA, Grünewald K, Koster AJ, Snijder EJ, Bárcena M. A molecular pore spans the double membrane of the coronavirus replication organelle. Science. 2020 Sep 11;369(6509):1395–8. DOI: 10.1126/science.abd3629

63. Stertz S, Reichelt M, Spiegel M, Kuri T, Martínez-Sobrido L, García-Sastre A, Weber F, Kochs G. The intracellular sites of early replication and budding of SARS-coronavirus. Virology. 2007 May;361(2):304–15. DOI: 10.1016/j.virol.2006.11.027

64. Guan W jie, Ni Z yi, Hu YY hua, Liang W hua, Ou C quan, He J xing, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. New England Journal of Medicine. 2020 Apr 30 [cited 2021 Oct 20];382(18):1708–20. DOI: 10.1056/NEJMoa2002032

65. Wang Y, Qiao F, Zhou F, Yuan Y. Surface distribution of severe acute respiratory syndrome coronavirus 2 in Leishenshan Hospital in China. Indoor and Built Environment. 2020 Jul 28;26(7):1420326X2094293. DOI: 10.1177/1420326X20942938

66. Jayaweera M, Perera H, Gunawardana B, Manatunge J. Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. Environmental Research. 2020 Sep;188(January):109819. DOI: 10.1016/j.envres.2020.109819

67. Sender R, Bar-On YM, Gleizer S, Bernshtein B, Flamholz A, Phillips R, Milo R. The total number and mass of SARS-CoV-2 virions. Proceedings of the National Academy of Sciences. 2021 Jun 22;118(25):1–9. DOI: 10.1073/pnas.2024815118

68. Prentiss M, Chu A, Berggren KK. Superspreading Events Without Superspreaders: Using High Attack Rate Events to Estimate N . medRxiv. 2020 Oct 23 [cited 2022 Apr 29];2020.10.21.20216895. DOI: 10.1101/2020.10.21.20216895

69. World Health Organization(WHO). Infection prevention and control of epidemic- and pandemic-prone acute respiratory infections in health care. WHO Guidelines. 2014;1–156.

70. Siegel JD, Rhinehart E, Jackson M, Chiarello L, and the Healthcare Infection Control Practices Advisory Committee. 2007 Guideline for Isolation Precautions: Preventing Transmission of Infectious Agents in Healthcare Settings (updated July 2019). Centers for Diasease Control and Prevention. 2019;1–232.

71. Wang Y, Xu G, Huang YW. Modeling the load of SARS-CoV-2 virus in human expelled particles during coughing and speaking. PLoS ONE. 2020;15(10 October):1–12. DOI: 10.1371/journal.pone.0241539

72. van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, Tamin A, Harcourt JL, Thornburg NJ, Gerber SI, Lloyd-Smith JO, de Wit E, Munster VJ. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. New England Journal of Medicine. 2020 Apr 16;382(16):1564–7. DOI: 10.1056/NEJMc2004973

73. Liu Y, Li T, Deng Y, Liu S, Zhang D, Li H, Wang X, Jia L, Han J, Bei Z, Li L, Li J. Stability of SARS-CoV-2 on environmental surfaces and in human excreta. The Journal of hospital infection. 2021 Jan;107(January):105–7. DOI: 10.1016/j.jhin.2020.10.021

74. Salian VS, Wright JA, Vedell PT, Nair S, Li C, Kandimalla M, Tang X, Carmona Porquera EM, Kalari KR, Kandimalla KK. COVID-19 Transmission, Current Treatment, and Future Therapeutic Strategies. Molecular Pharmaceutics. 2021 Mar 1;18(3):754–71. DOI: 10.1021/acs.molpharmaceut.0c00608

75. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020 Feb;395(10223):497–506. DOI: 10.1016/S0140-6736(20)30183-5

76. Duzgun SA, Durhan G, Demirkazik FB, Akpinar MG, Ariyurek OM. COVID-19 pneumonia: the great radiological mimicker. Insights into Imaging. 2020;11(1). DOI: 10.1186/s13244-020-00933-z

77. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China. JAMA. 2020 Apr 7;323(13):1239. DOI: 10.1001/jama.2020.2648

78. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020 Mar 17;323(11):1061. DOI: 10.1001/jama.2020.1585

79. Zhang X, Li S, Niu S. ACE2 and COVID-19 and the resulting ARDS. Postgraduate Medical Journal. 2020 Jul;96(1137):403–7. DOI: 10.1136/postgradmedj-2020-137935

80. Soy M, Keser G, Atagündüz P, Tabak F, Atagündüz I, Kayhan S. Cytokine storm in COVID-19: pathogenesis and overview of anti-inflammatory agents used in treatment. Clinical Rheumatology. 2020 Jul 30;39(7):2085–94. DOI: 10.1007/s10067-020-05190-5

81. Gao YL, Du Y, Zhang C, Cheng C, Yang HY, Jin YF, Duan GC, Chen SY. Role of renin-angiotensin system in acute lung injury caused by viral infection. Vol. 13, Infection and Drug Resistance. 2020. p. 3715–25. DOI: 10.2147/IDR.S265718

82. Wishart DS. Metabolomics for Investigating Physiological and Pathophysiological Processes. Physiological Reviews. 2019 Oct 1;99(4):1819–75. DOI: 10.1152/physrev.00035.2018

83. Muthubharathi BC, Gowripriya T, Balamurugan K. Metabolomics: small molecules that matter more. Molecular Omics. 2021;17(2):210–29. DOI: 10.1039/D0MO00176G

84. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R. Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics. 2007 Sep;8(9):1243–66. DOI: 10.2217/14622416.8.9.1243

85. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome

Database. Nucleic Acids Research. 2007 Jan 3;35(Database):D521–6. DOI: 10.1093/nar/gkl923

86. Kell DB, Oliver SG. The metabolome 18 years on: a concept comes of age. Metabolomics. 2016 Sep 2;12(9):148. DOI: 10.1007/s11306-016-1108-4

87. Li X, Fan X, Yang H, Liu Y. Review of Metabolomics-Based Biomarker Research for Parkinson's Disease. Molecular Neurobiology. 2022 Feb 26;59(2):1041–57. DOI: 10.1007/s12035-021-02657-7

88. Lotta LA, Scott RA, Sharp SJ, Burgess S, Luan J, Tillin T, et al. Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. PLOS Medicine. 2016 Nov 29;13(11):e1002179. DOI: 10.1371/journal.pmed.1002179

89. Hrydziuszko O, Perera MTPR, Laing R, Kirwan J, Silva MA, Richards DA, Murphy N, Mirza DF, Viant MR. Mass Spectrometry Based Metabolomics Comparison of Liver Grafts from Donors after Circulatory Death (DCD) and Donors after Brain Death (DBD) Used in Human Orthotopic Liver Transplantation. PLOS ONE. 2016 Nov 11;11(11):e0165884. DOI: 10.1371/journal.pone.0165884

90. Nieman DC, Shanely RA, Gillitt ND, Pappan KL, Lila MA. Serum Metabolic Signatures Induced By a Three-Day Intensified Exercise Period Persist After 14 h of Recovery in Runners. Journal of Proteome Research. 2013 Oct 4;12(10):4577–84. DOI: 10.1021/pr400717j

91. Jacob M, Lopata AL, Dasouki M, Abdel Rahman AM. Metabolomics toward personalized medicine. Mass Spectrometry Reviews. 2019 May 26;38(3):221–38. DOI: 10.1002/mas.21548

92. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research. 2018 Jan 4;46(D1):D608–17. DOI: 10.1093/nar/gkx1089

93. Attard JA, Dunn WB, Mergental H, Mirza DF, Afford SC, Perera MTPR. Systematic Review: Clinical Metabolomics to Forecast Outcomes in Liver Transplantation Surgery. OMICS: A Journal of Integrative Biology. 2019 Oct 1;23(10):463–76. DOI: 10.1089/omi.2019.0086

94. Alseekh S, Aharoni A, Brotman Y, Contrepois K, D'Auria J, Ewald J, et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nature Methods. 2021 Jul 8;18(7):747–56. DOI: 10.1038/s41592-021-01197-1

95. Vinayavekhin N, Saghatelian A. Untargeted Metabolomics. InJohn Wiley & Sons, Inc.John Wiley & Sons, Inc. John Wiley & Sons, Inc.John Wiley & Sons, Inc. 2010. p. 1–24. DOI: 10.1002/0471142727.mb3001s90

96. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, Dunn WB. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics. 2018 Jun 18;14(6):72. DOI: 10.1007/s11306-018-1367-3

97. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. Journal of the American Society for Mass Spectrometry. 2016 Dec 1;27(12):1897–905. DOI: 10.1007/s13361-016-1469-y

98. Wong A, Li X, Lau S, Woo P. Global Epidemiology of Bat Coronaviruses. Viruses. 2019 Feb 20;11(2):174. DOI: 10.3390/v11020174

99. Ghinai I, McPherson TD, Hunter JC, Kirking HL, Christiansen D, Joshi K, et al. First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. Lancet (London, England). 2020;395(10230):1137–44. DOI: 10.1016/S0140-6736(20)30607-3

100. Roberts I, Muelas MW, Taylor JM, Davison AS, Xu Y, Grixti JM, Gotts N, Sorokin A, Goodacre R, Kell DB. Untargeted metabolomics of COVID-19 patient serum reveals potential prognostic markers of both severity and outcome. medRxiv. 2020;792:2020.12.09.20246389. DOI: 10.1101/2020.12.09.20246389

101. Doğan HO, Şenol O, Bolat S, Yıldız ŞN, Büyüktuna SA, Sarısmailoğlu R, Doğan K, Hasbek M, Hekim SN. Understanding the pathophysiological changes via untargeted metabolomics in COVID-19 patients. Journal of Medical Virology. 2021 Apr 17;93(4):2340–9. DOI: 10.1002/jmv.26716

102. Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and Metabolomic Characterization of COVID-

19 Patient Sera. Cell. 2020 Jul;182(1):59-72.e15. DOI: 10.1016/j.cell.2020.05.032

103. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell. 2020 Dec;183(6):1479-1495.e20. DOI: 10.1016/j.cell.2020.10.037

104. Thomas T, Stefanoni D, Reisz JA, Nemkov T, Bertolone L, Francis RO, Hudson KE, Zimring JC, Hansen KC, Hod EA, Spitalnik SL, D'Alessandro A. COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status. JCI Insight. 2020 Jul 23;5(14). DOI: 10.1172/jci.insight.140327

105. Páez-Franco JC, Torres-Ruiz J, Sosa-Hernández VA, Cervantes-Díaz R, Romero-Ramírez S, Pérez-Fragoso A, et al. Metabolomics analysis reveals a modified amino acid metabolism that correlates with altered oxygen homeostasis in COVID-19 patients. Scientific Reports. 2021;11(1):1–12. DOI: 10.1038/s41598-021-85788-0

106. Atkins PW (Peter W, Jones L. Chemical principles : the quest for insight. 2010;

107. L. G. Wade J. Infrared Spectroscoy and Mass Spectrometry. In Pearson 2011. p. 513–62.

108. Wood BR. The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues. Chemical Society Reviews. 2016;45(7):1980–98. DOI: 10.1039/C5CS00511F

109. Movasaghi Z, Rehman S, ur Rehman DI. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. Applied Spectroscopy Reviews. 2008 Feb;43(2):134–79. DOI: 10.1080/05704920701829043

110. Cakmak G, Zorlu F, Severcan M, Severcan F. Screening of protective effect of amifostine on radiation-induced structural and functional variations in rat liver microsomal membranes by FT-IR spectroscopy. Analytical chemistry. 2011 Apr 1;83(7):2438–44. DOI: 10.1021/ac102043p

111. R T, GJ P, D R, G S. XCMS Online: a web-based platform to process untargeted metabolomic data. Analytical chemistry. 2012 Jun 5 [cited 2021 Oct 11];84(11):5035–9. DOI: 10.1021/AC300698C

112. Patti GJ, Tautenhahn R, Rinehart D, Cho K, Shriver LP, Manchester M, Nikolskiy I, Johnson CH, Mahieu NG, Siuzdak G. A View from Above: Cloud Plots to Visualize Global Metabolomic Data. Analytical Chemistry. 2013 Jan 15;85(2):798–804. DOI: 10.1021/ac3029745

113. Gowda H, Ivanisevic J, Johnson CH, Kurczy ME, Benton HP, Rinehart D, et al. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. Analytical Chemistry. 2014 Jul 15;86(14):6931–9. DOI: 10.1021/ac500734c

114. Zhu ZJ, Schultz AW, Wang J, Johnson CH, Yannone SM, Patti GJ, Siuzdak G. Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. Nature Protocols. 2013 Mar 7;8(3):451–60. DOI: 10.1038/nprot.2013.004

115. Pluskal T, Korf A, Smirnov A, Schmid R, Fallon TR, Du X, Weng JK. CHAPTER 7. Metabolomics Data Analysis Using MZmine. InThe Royal Society of ChemistryThe Royal Society of Chemistry The Royal Society of ChemistryThe Royal Society of Chemistry 2020. p. 232–54. DOI: 10.1039/9781788019880-00232

116. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology. 2012 Oct;30(10):918–20. DOI: 10.1038/nbt.2377

117. Total Ion Chromatogram (TIC) : SHIMADZU (Shimadzu Corporation). [cited 2022 May 4].

118. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 Sep;20(3):273–97. DOI: 10.1007/BF00994018

119. Bishop CM. Neural networks for pattern recognition. Clarendon Press ;;Oxford University PressClarendon Press ;;Oxford University Press Clarendon Press ;;Oxford University PressClarendon Press ;;Oxford University Press 1995. 482 p.

120. Wyner AJ, Olson M, Bleich J, Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. Journal of Machine Learning Research. 2015 Apr 28;18:1–33.

121. Atallah DM, Badawy M, El-Sayed A, Ghoneim MA. Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier. Multimedia Tools and Applications. 2019 Jul 28;78(14):20383–407. DOI: 10.1007/s11042-019-7370-5

122. Ketkar N. Stochastic Gradient Descent. InApressApress ApressApress 2017. p. 113–32. DOI: 10.1007/978-1-4842-2766-4_8

123. DOGAN A, GURBANOV R, SEVERCAN M, SEVERCAN F. CoronaVac (Sinovac) COVID-19 vaccine-induced molecular changes in healthy human serum by infrared spectroscopy coupled with chemometrics. Turkish Journal of Biology. 2021 Aug 30;45(SI-1):549–58. DOI: 10.3906/biy-2105-65

124. Zhang B, Zhou X, Zhu C, Song Y, Feng F, Qiu Y, Feng J, Jia Q, Song Q, Zhu B, Wang J. Immune Phenotyping Based on the Neutrophil-to-Lymphocyte Ratio and IgG Level Predicts Disease Severity and Outcome for Patients With COVID-19. Frontiers in Molecular Biosciences. 2020 Jul 3;7(July):1–7. DOI: 10.3389/fmolb.2020.00157

125. Huang Y, Lu Y, Huang YM, Wang M, Ling W, Sui Y, Zhao HL. Obesity in patients with COVID-19: a systematic review and meta-analysis. Metabolism: Clinical and Experimental. 2020 Dec;113(January):154378. DOI: 10.1016/j.metabol.2020.154378

126. Cornwell CR, Hsu J, Tompkins LK, Pennington AF, Flanders WD, Sircar K. Clinical outcomes among hospitalized US adults with asthma or chronic obstructive pulmonary disease, with or without COVID-19. Journal of Asthma. 2021 [cited 2022 May 22]; DOI: 10.1080/02770903.2021.2018703

127. Bandeira CCS, Madureira KCR, Rossi MB, Gallo JF, da Silva APMA, Torres VL, de Lima VA, Júnior NK, Almeida JD, Zerbinati RM, Braz-Silva PH, Lindoso JAL, da Silva Martinho H. Micro-Fourier-transform infrared reflectance spectroscopy as tool for probing IgG glycosylation in COVID-19 patients. Scientific Reports. 2022;12(1):1–13. DOI: 10.1038/s41598-022-08156-6

128. Nogueira MS, Leal LB, Macarini W, Pimentel RL, Muller M, Vassallo PF, Campos LCG, dos Santos L, Luiz WB, Mill JG, Barauna VG, de Carvalho LF das C e. S. Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. Scientific Reports. 2021;11(1):1–13. DOI: 10.1038/s41598-021-93511-2

129. Wishart DS, Guo AC, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the Human Metabolome Database for 2022. Nucleic acids research. 2022 Jan 7 [cited 2022 May 26];50(D1):D622–31. DOI: 10.1093/NAR/GKAB1062

130. Xia J, Wishart DS. Metabolomic Data Processing, Analysis, and Interpretation Using MetaboAnalyst. Current Protocols in Bioinformatics. 2011 Jun;34(1):1–48. DOI: 10.1002/0471250953.bi1410s34

131. Vollbracht C, Kraft K. Oxidative Stress and Hyper-Inflammation as Major Drivers of Severe COVID-19 and Long COVID: Implications for the Benefit of High-Dose Intravenous Vitamin C. Frontiers in Pharmacology. 2022 Apr 29;13(April):1–14. DOI: 10.3389/fphar.2022.899198

132. Rowland A, Miners JO, Mackenzie PI. The UDP-glucuronosyltransferases: Their role in drug metabolism and detoxification. The International Journal of Biochemistry & Cell Biology. 2013 Jun;45(6):1121–32. DOI: 10.1016/j.biocel.2013.02.019

133. Shakoor H, Feehan J, Mikkelsen K, Al Dhaheri AS, Ali HI, Platat C, Ismail LC, Stojanovska L, Apostolopoulos V. Be well: A potential role for vitamin B in COVID-19. Maturitas. 2021 Feb;144:108–11. DOI: 10.1016/j.maturitas.2020.08.007

134. Xu Y, Baylink DJ, Chen CS, Reeves ME, Xiao J, Lacy C, Lau E, Cao H. The importance of vitamin d metabolism as a potential prophylactic, immunoregulatory and neuroprotective treatment for COVID-19. Journal of Translational Medicine. 2020 Dec 26;18(1):322. DOI: 10.1186/s12967-020-02488-5

135. Kimura T, Yasuda K, Yamamoto R, Soga T, Rakugi H, Hayashi T, Isaka Y. Identification of biomarkers for development of end-stage kidney disease in chronic kidney disease by metabolomic profiling. Scientific Reports. 2016 May 18;6(1):26138. DOI: 10.1038/srep26138

136. Şehirli AÖ, Sayıner S. Daylight is critical to preserve 5-methoxytryptophol levels in suspected and confirmed COVID-19 patients. Medical Hypotheses. 2021 Feb;147(January):110504. DOI: 10.1016/j.mehy.2021.110504

137. Rucker RB, Bauerly K. Handbook of Vitamins. Handbook of Vitamins. CRC PressCRC Press CRC PressCRC Press 2007 [cited 2022 May 27]. DOI: 10.1201/9781420005806

138. Pereyra AS, Harris KL, Soepriatna AH, Waterbury QA, Bharathi SS, Zhang Y, Fisher-Wellman KH, Goergen CJ, Goetzman ES, Ellis JM. Octanoate is differentially metabolized in liver and muscle and fails to rescue cardiomyopathy in CPT2 deficiency. Journal of Lipid Research. 2021;62:100069. DOI: 10.1016/j.jlr.2021.100069

139. Lehmann M, Allers K, Heldt C, Meinhardt J, Schmidt F, Rodriguez-Sillke Y, et al. Human small intestinal infection by SARS-CoV-2 is characterized by a mucosal infiltration with activated CD8+ T cells. Mucosal Immunology. 2021;14(6):1381–92. DOI: 10.1038/s41385-021-00437-z

140. Wang Q, Fang P, He R, Li M, Yu H, Zhou L, et al. O -GlcNAc transferase promotes influenza A virus–induced cytokine storm by targeting interferon regulatory factor–5. Science Advances. 2020 Apr 17;6(16):1–14. DOI: 10.1126/sciadv.aaz7086

141. Chen X, Zhou L, Peng N, Yu H, Li M, Cao Z, et al. MicroRNA-302a suppresses influenza A virus–stimulated interferon regulatory factor-5 expression and cytokine storm induction. Journal of Biological Chemistry. 2017 Dec;292(52):21291–303. DOI: 10.1074/jbc.M117.805937

142. Braz-de-Melo HA, Faria SS, Pasquarelli-do-Nascimento G, Santos I de O, Kobinger GP, Magalhães KG. The Use of the Anticoagulant Heparin and Corticosteroid Dexamethasone as Prominent Treatments for COVID-19. Frontiers in Medicine. 2021 Apr 23;8(April):1–11. DOI: 10.3389/fmed.2021.615333

143. Shi C, Tingting W, Li JP, Sullivan MA, Wang C, Wang H, Deng B, Zhang Y. Comprehensive Landscape of Heparin Therapy for COVID-19. Carbohydrate Polymers. 2021;254(August 2020):117232. DOI: 10.1016/j.carbpol.2020.117232

144. Enns GM, Berry SA, Berry GT, Rhead WJ, Brusilow SW, Hamosh A. Survival after Treatment with Phenylacetate and Benzoate for Urea-Cycle Disorders. New England Journal of Medicine. 2007 May 31;356(22):2282–92. DOI: 10.1056/NEJMoa066596

145. Dator R, Villalta PW, Thomson N, Jensen J, Hatsukami DK, Stepanov I, Warth B, Balbo S. Metabolomics Profiles of Smokers from Two Ethnic Groups with Differing Lung Cancer Risk. Chemical Research in Toxicology. 2020 Aug 17;33(8):2087–98. DOI: 10.1021/acs.chemrestox.0c00064

146. Poupore N, Chosed R, Arce S, Rainer R, Goodwin RL, Nathaniel TI. Metabolomic Profiles of Men and Women Ischemic Stroke Patients. Diagnostics. 2021 Sep 28;11(10):1786. DOI: 10.3390/diagnostics11101786

Appendix

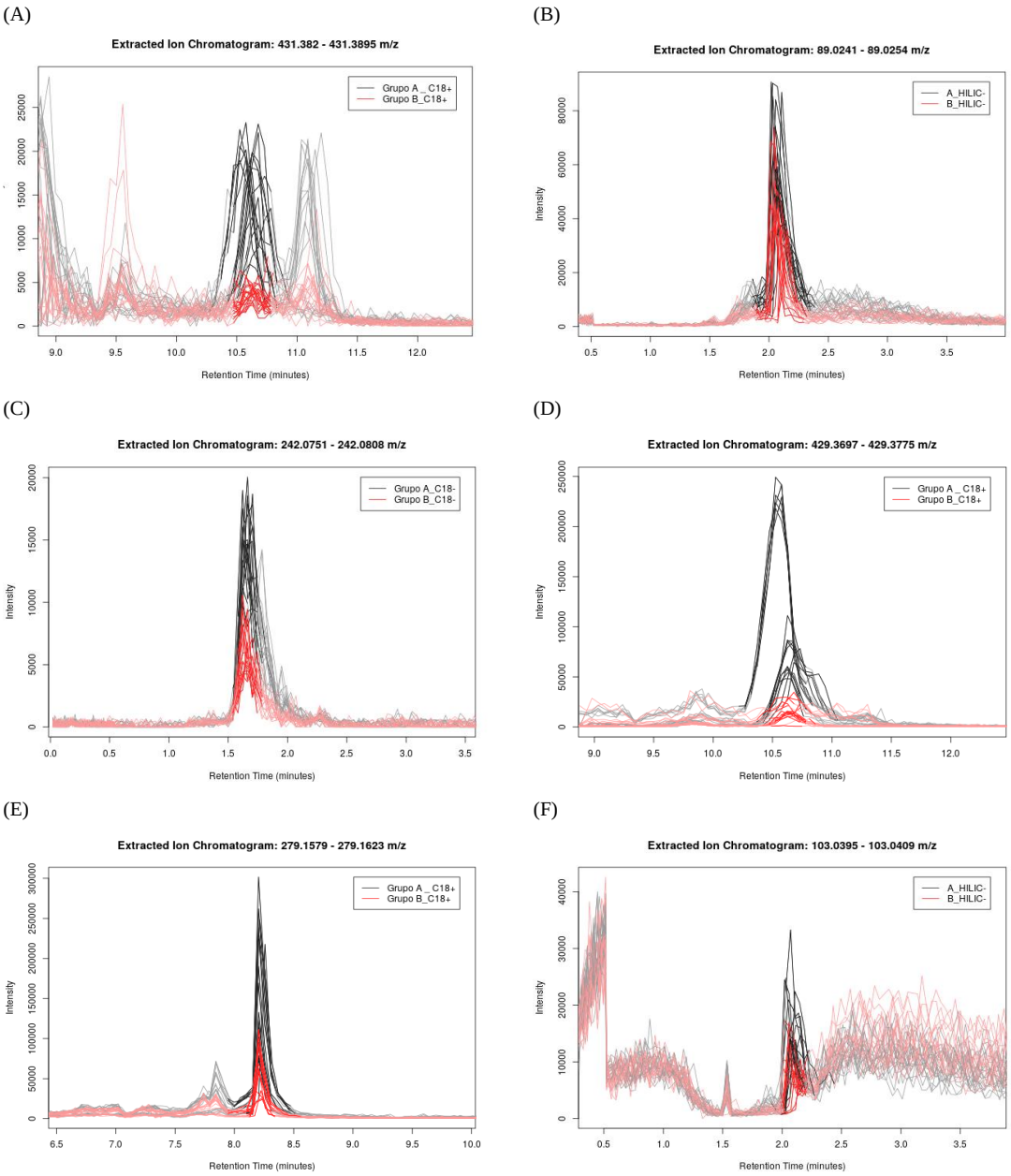
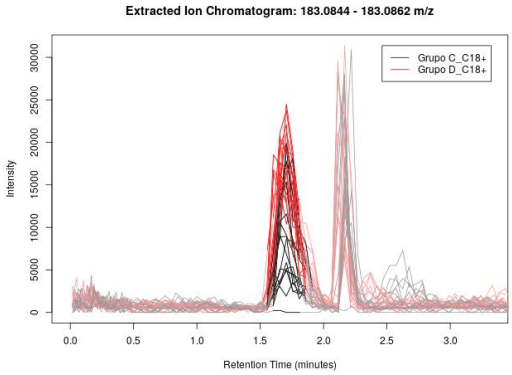
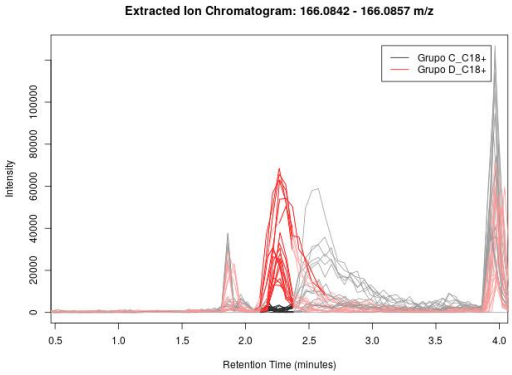


Figure 1 – 1st assay EIC's of (A) α-tocopherol; (B) (R)-lactate; (C) Cytidine; (D) 4α-hydroxymethyl-4β-methyl-5α-zymosterol; (E) α-carboxyethylhydroxychroman; (F) (R)-3-hydroxybutanoate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group B (red).

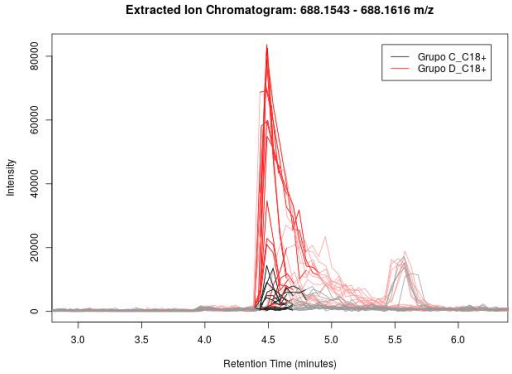
(A)



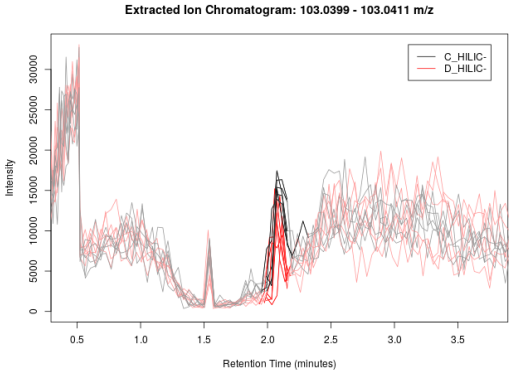
(B)



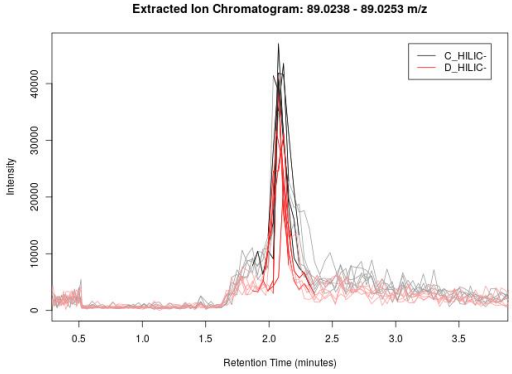
(C)



(D)



(E)



(F)

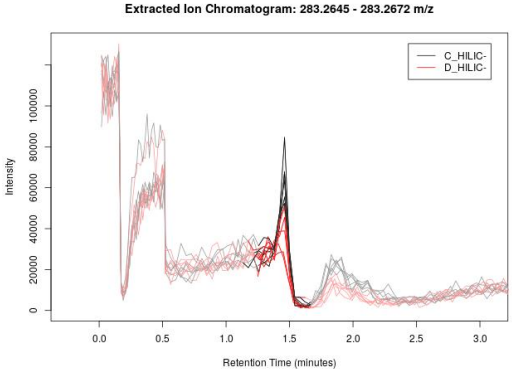
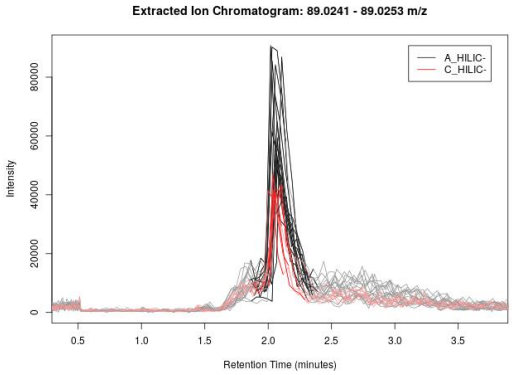
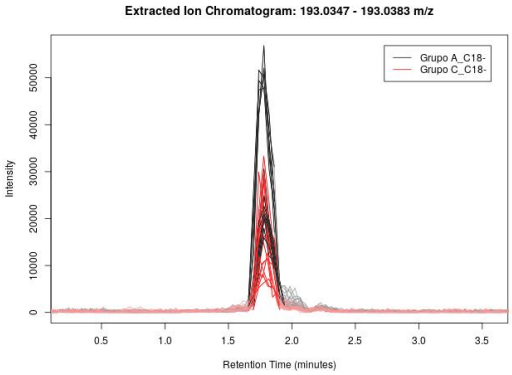


Figure 3- 1st assay EIC's of (A) D-sorbitol; (B) 4-(3-pyridyl)-butanoate; (C) 3'-dephospho-CoA; (D) (R)-3-hydroxybutanoate; (E) (R)-lactate; (F) stearate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group C (Black) and Group D (red).

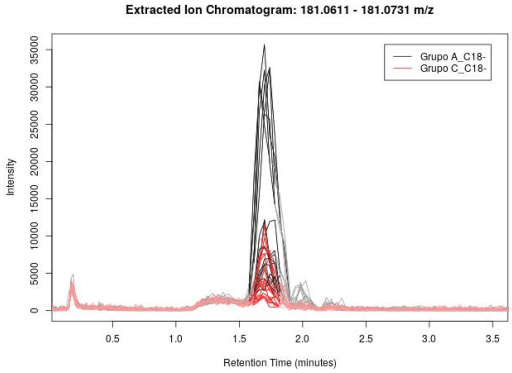
(A)


Extracted Ion Chromatogram: 89.0241 - 89.0253 m/z
A_HILIC-
C_HILIC-
Intensity
Retention Time (minutes)

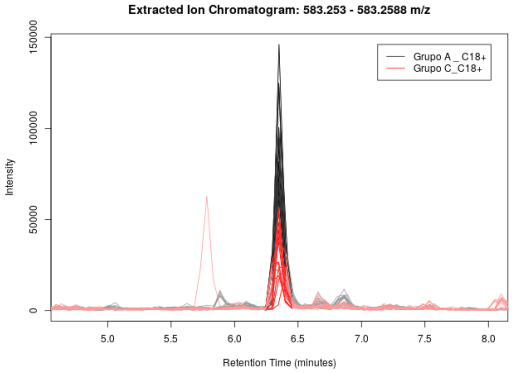
(B)


Extracted Ion Chromatogram: 193.0347 - 193.0383 m/z
Grupo A_C18-
Grupo C_C18-
Intensity
Retention Time (minutes)

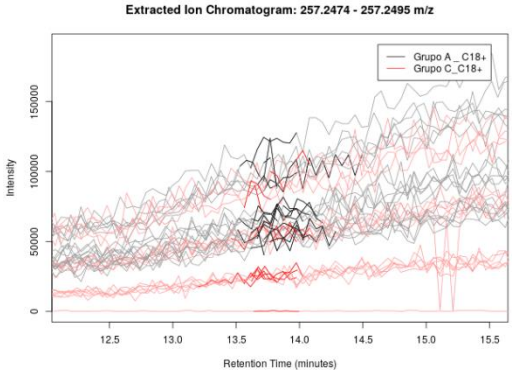
(C)


Extracted Ion Chromatogram: 181.0611 - 181.0731 m/z
Grupo A_C18-
Grupo C_C18-
Intensity
Retention Time (minutes)

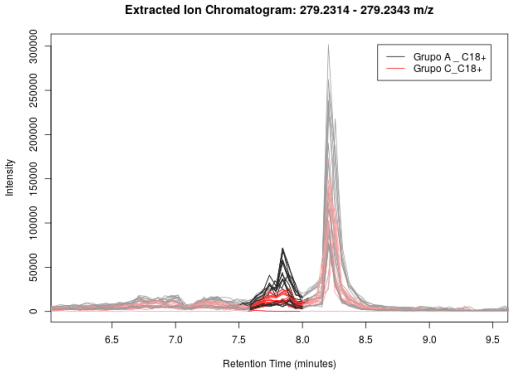
(D)


Extracted Ion Chromatogram: 583.253 - 583.2588 m/z
Grupo A _C18+
Grupo C_C18+
Intensity
Retention Time (minutes)

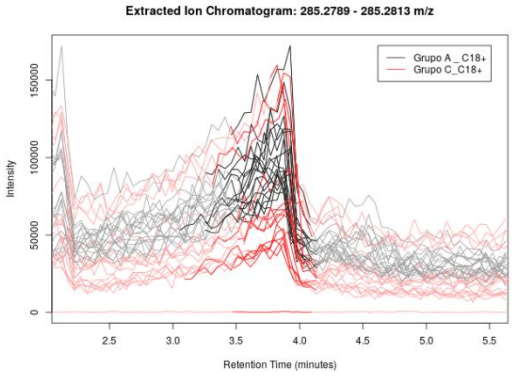
(E)


Extracted Ion Chromatogram: 257.2474 - 257.2495 m/z
Grupo A _C18+
Grupo C_C18+
Intensity
Retention Time (minutes)

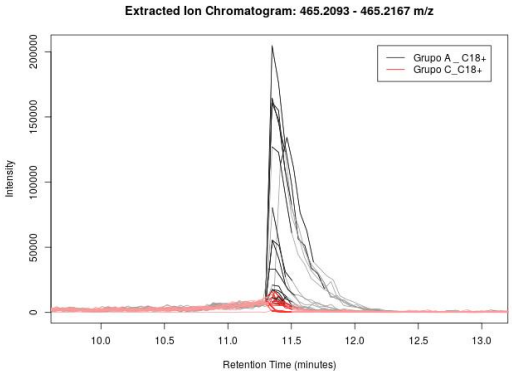
(F)


Extracted Ion Chromatogram: 279.2314 - 279.2343 m/z
Grupo A _C18+
Grupo C_C18+
Intensity
Retention Time (minutes)

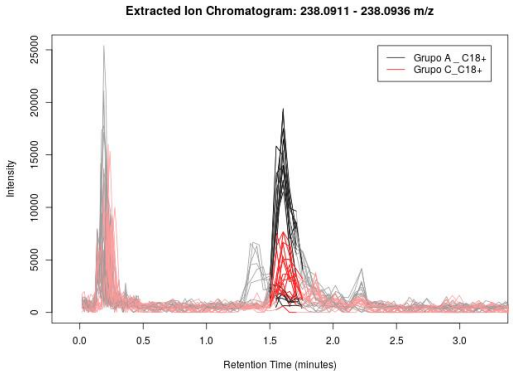
(G)


Extracted Ion Chromatogram: 285.2789 - 285.2813 m/z
Grupo A _C18+
Grupo C_C18+
Intensity
Retention Time (minutes)

(H)


Extracted Ion Chromatogram: 465.2093 - 465.2167 m/z
Grupo A _C18+
Grupo C_C18+
Intensity
Retention Time (minutes)

(I)



(J)

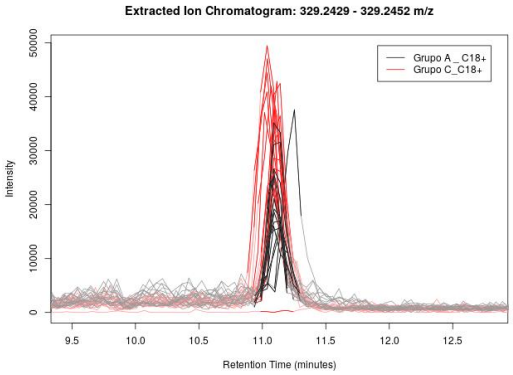
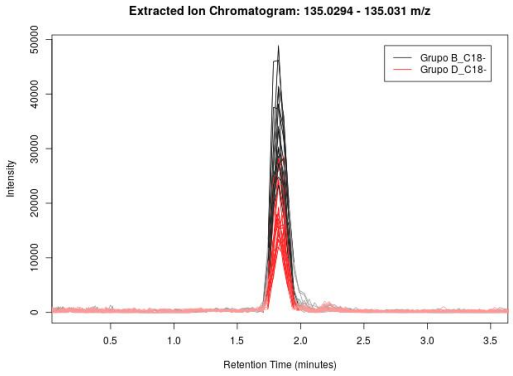
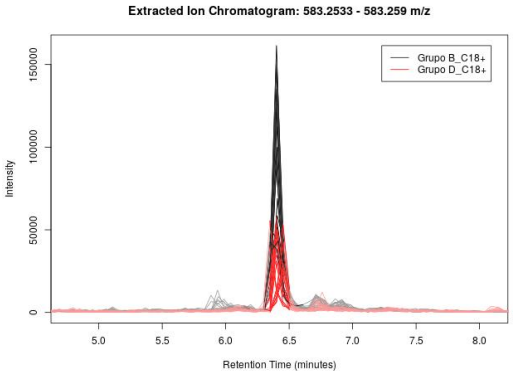


Figure 4 – 1st assay EIC's of (A) α-D-glucose; (B) α-L-iduronate; (C) D-sorbitol; (D) biliverdin-IX-α; (E) palmitate; (F) α-linolenate; (G) stearate; (H) 4-hydroxy-2-nonenal-glutathione conjugate; (I) 6-pyruvoyl tetrahydropterin; (J) (4Z,7Z,10Z,13Z,16Z,19Z)-docosahexaenoate; (K) 4-(3-pyridyl)-butanoate; (L) N-acetyl-β-neuraminate; (M) 2-trans-hexadecenal. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group C (red).

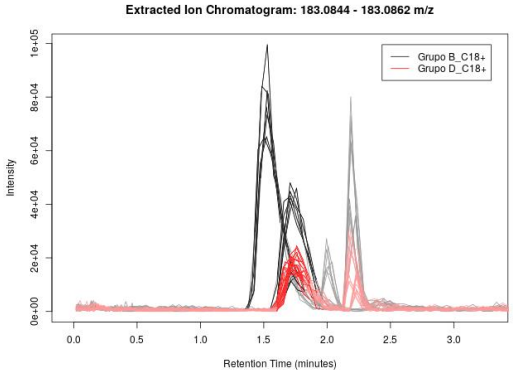
(A)



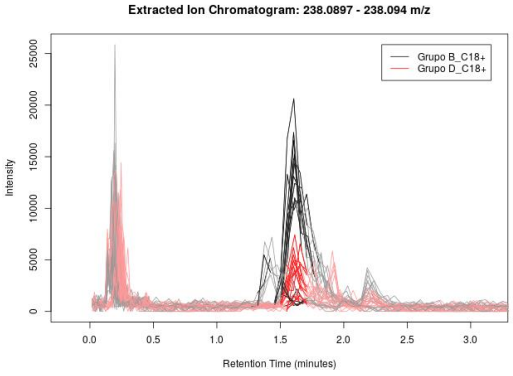
(B)



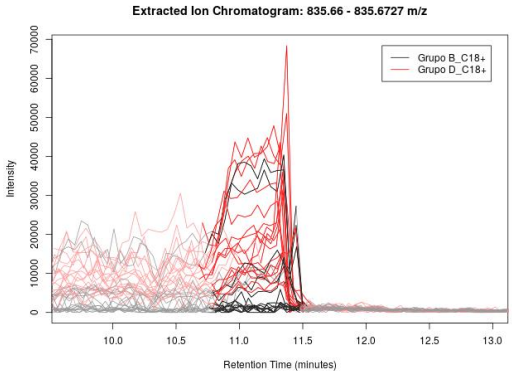
(C)



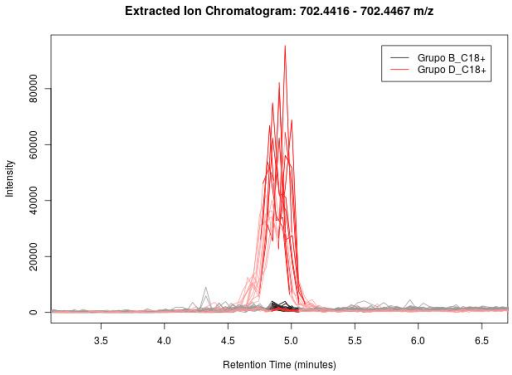
(D)



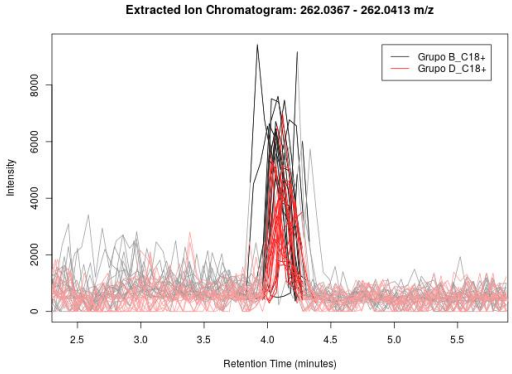
(E)



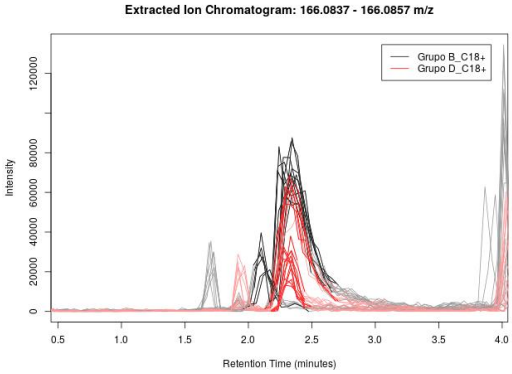
(F)



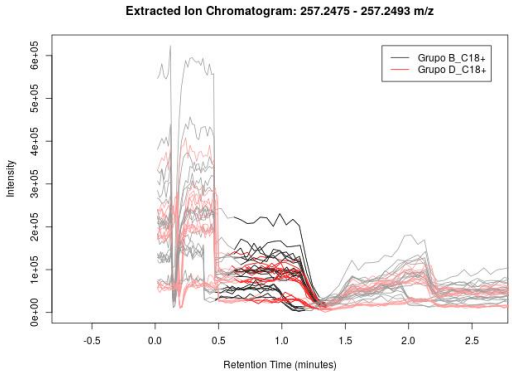
(G)



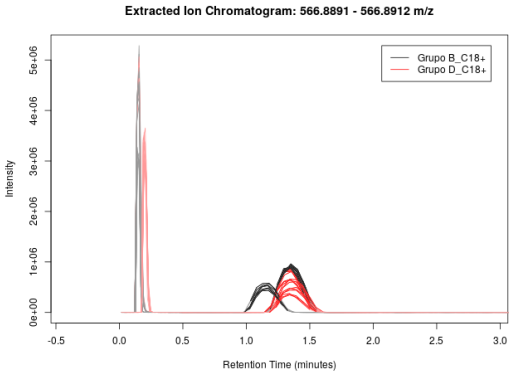
(H)



(I)



(J)



(K)

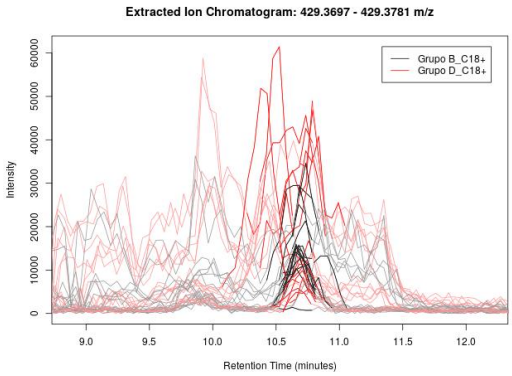
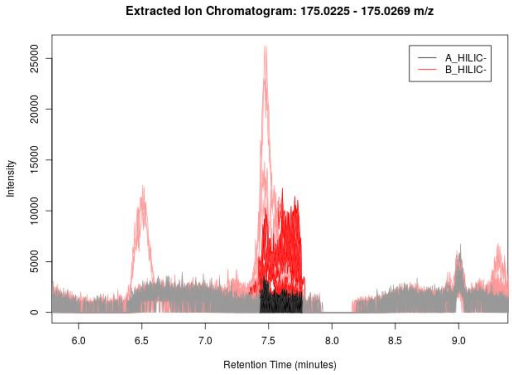
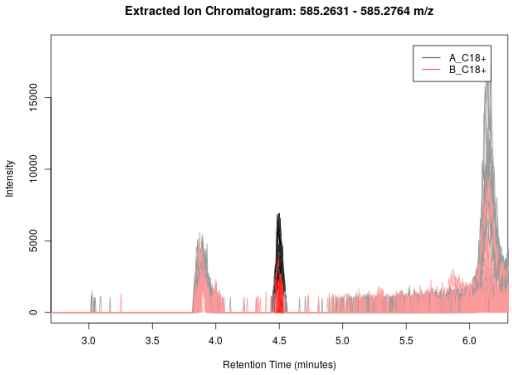


Figure 5 - 1st assay EIC's of (A) hypoxanthine; (B) biliverdin-IX-α; (C) D-sorbitol; (D) N-formylkynurenine; (E) 3,4-dihydroxy-5-all-trans-decaprenylbenzoate; (F) sulfate; (G) α-L-iduronate; (H) 4-(3-pyridyl)-butanoate; (I) palmitate; (J) N-acetyl-β-neuraminate; (K) 4α-hydroxymethyl-4β-methyl-5α-zymosterol. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group B (Black) and Group D (red).

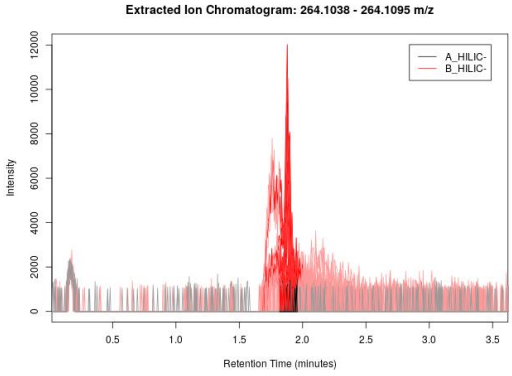
(A)



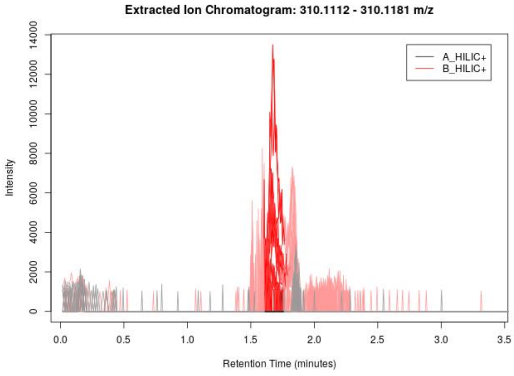
(B)



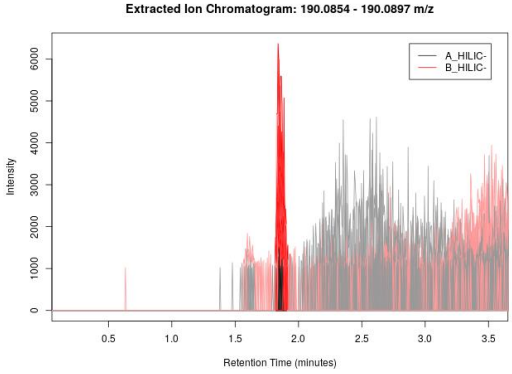
(C)



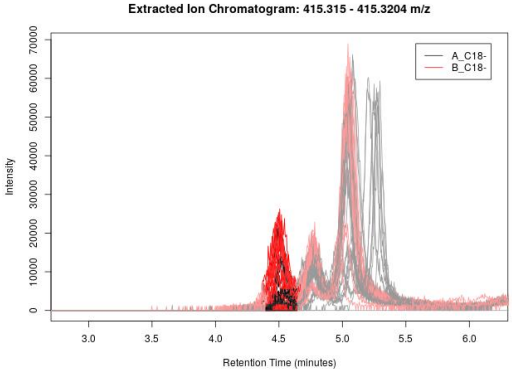
(D)



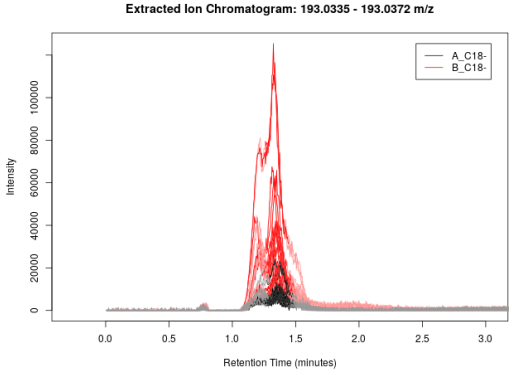
(E)



(F)



(G)



(H)

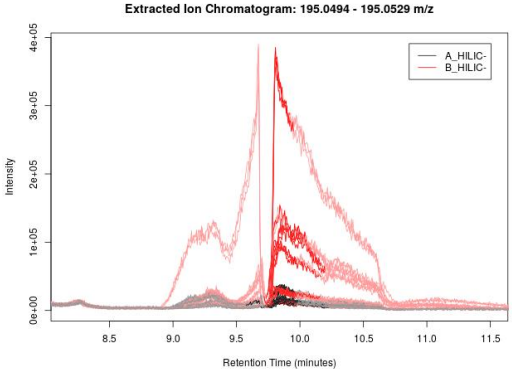
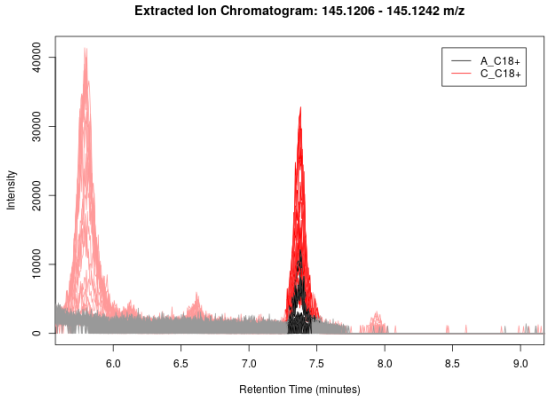
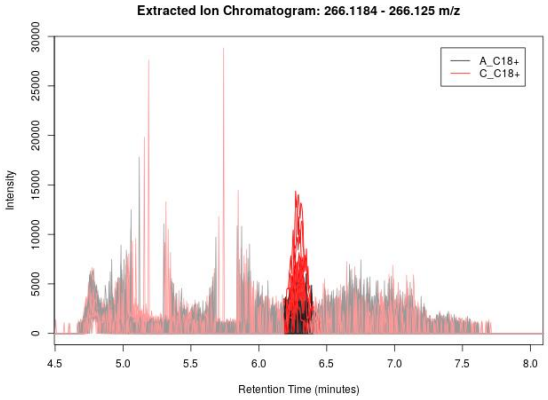


Figure 6 - 2nd assay EIC's of (A) L-ascorbate; (B) bilirubin; (C) thiamin; (D) N-acetyl-β-neuraminate; (E) 5-methoxytryptophol; (F) calcitriol; (G) aldehydo-D-glucuronate; (H) L-gulonate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group B (red)

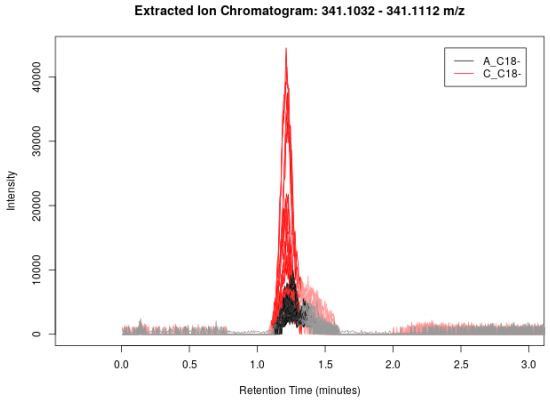
(A)


Extracted Ion Chromatogram: 145.1206 - 145.1242 m/z
A_C18+
C_C18+
Intensity
Retention Time (minutes)

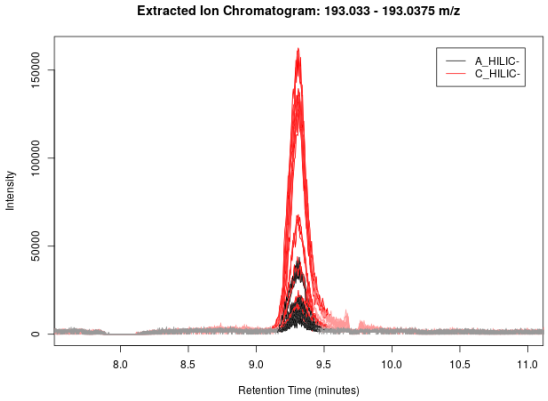
(B)


Extracted Ion Chromatogram: 266.1184 - 266.125 m/z
A_C18+
C_C18+
Intensity
Retention Time (minutes)

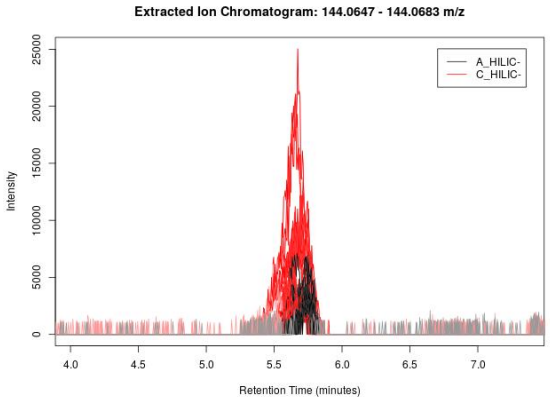
(C)


Extracted Ion Chromatogram: 341.1032 - 341.1112 m/z
A_C18-
C_C18-
Intensity
Retention Time (minutes)

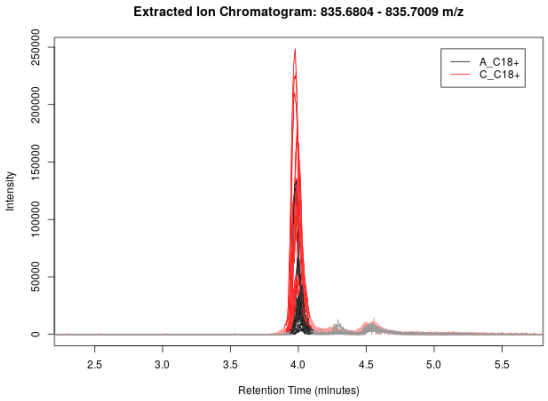
(D)


Extracted Ion Chromatogram: 193.033 - 193.0375 m/z
A_HILIC-
C_HILIC-
Intensity
Retention Time (minutes)

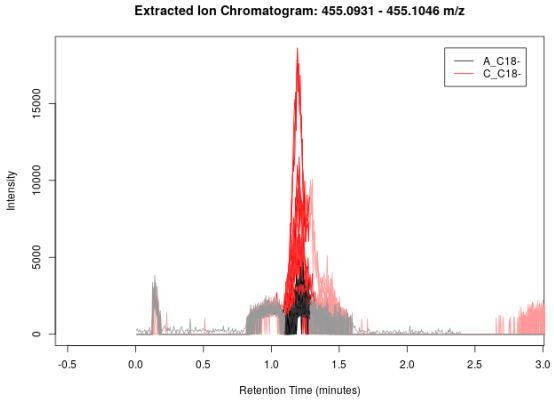
(E)


Extracted Ion Chromatogram: 144.0647 - 144.0683 m/z
A_HILIC-
C_HILIC-
Intensity
Retention Time (minutes)

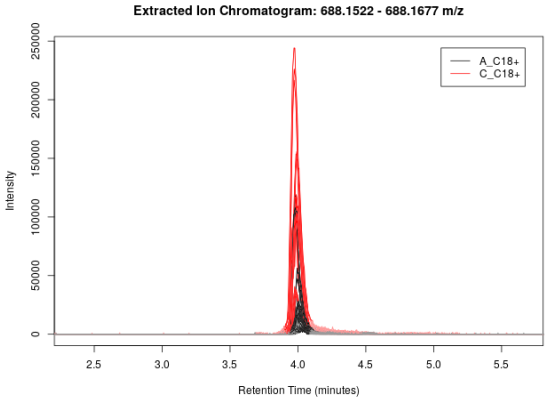
(F)


Extracted Ion Chromatogram: 835.6804 - 835.7009 m/z
A_C18+
C_C18+
Intensity
Retention Time (minutes)

(G)


Extracted Ion Chromatogram: 455.0931 - 455.1046 m/z
A_C18-
C_C18-
Intensity
Retention Time (minutes)

(H)


Extracted Ion Chromatogram: 688.1522 - 688.1677 m/z
A_C18+
C_C18+
Intensity
Retention Time (minutes)

(I)

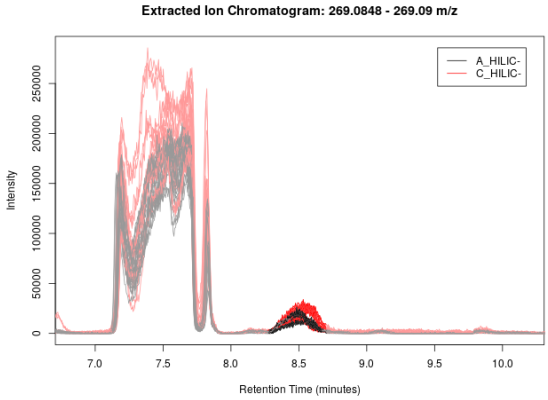


Figure 7 - 2nd assay EIC's of: (A) Octanoate; (B) Thiamin; (C) Sucrose; (D) α-L-iduronate; (E) 4-acetamidobutanoate;(F) 6-methoxy-3-methyl-2-decaprenyl-1,4-benzoquinol; (G) FMN; (H) 3'-dephospho-CoA; (I) α,α-trehalose. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group A (Black) and Group C (red).

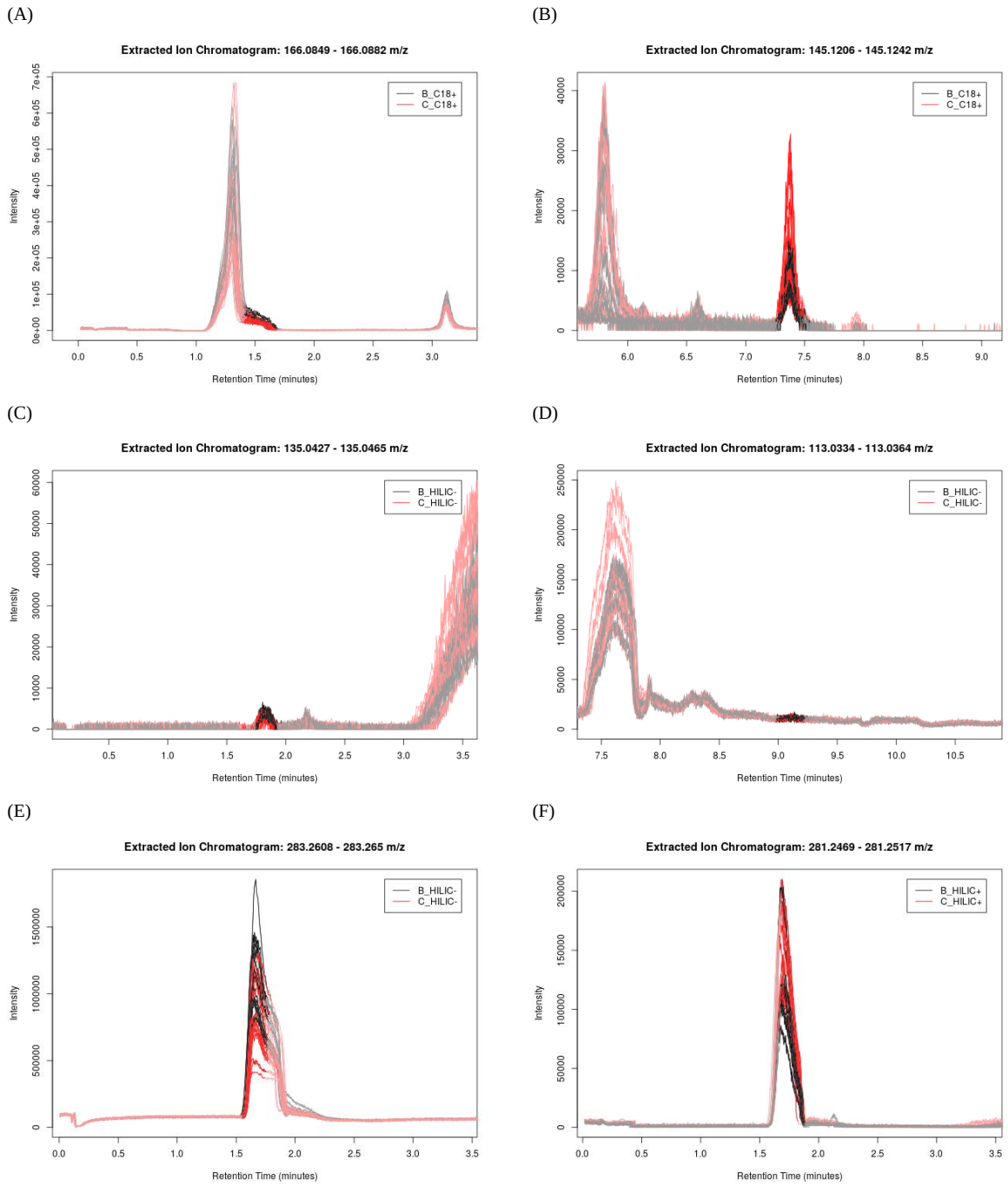


Figure 8 - 2nd assay EIC's of: (A) 4-(3-pyridyl)-butanoate; (B) octanoate; (C) phenylacetate; (D) 5,6-dihydrouracil; (E) stearate; (F) linoleate. Chromatograms are represented with intensity for the given m/z in Y axis with the retention time for that same m/z in X axis. Signals are separated in Group B (Black) and Group C (red).