Universidade de Lisboa

Faculdade de Farmácia

FTIR-ATR analysis of CoVid-19 blood plasma for immunological biochemical indicators estimation

Alexandra Raquel Moreira Martins

Trabalho de Campo orientado pelo Professor Doutor João Pedro Martins de Almeida Lopes, Professor Auxiliar

Mestrado Integrado em Ciências Farmacêuticas

2022

Universidade de Lisboa

Faculdade de Farmácia

FTIR-ATR analysis of CoVid-19 blood plasma for immunological biochemical indicators estimation

Alexandra Raquel Moreira Martins

Trabalho Final de Mestrado Integrado em Ciências Farmacêuticas apresentado à Universidade de Lisboa através da Faculdade de Farmácia

Trabalho de Campo orientado pelo Professor Doutor João Pedro Martins de Almeida Lopes, Professor Auxiliar

2022

Resumo

Neste estudo, avaliamos a exatidão dos resultados, ao usar a Espectroscopia de Infravermelho com Transformada de Fourier com Reflexão Total Atenuada, juntamente com análise multivariada, para analisar e identificar amostras de plasma de pacientes positivos e negativos para a CoVid-19.

O processo de triagem de CoVid-19 tornou-se um processo difícil, devido ao aumento de casos e, consequentemente, a afluência aos hospitais. Com isto, a hipótese de utilizar esta técnica como ferramenta diagnóstica, poderia ser um método mais rápido para identificar pacientes com indicadores inflamatórios mais elevados.

Primeiramente, foi realizada uma análise exploratória utilizando a Análise de Componentes Principais, utilizando os espectros de amostras de pacientes de diversas origens, vacinados com as vacinas Pfizer/BioNTech e AstraZeneca.

Numa fase seguinte, foi desenvolvida uma Análise Discriminante de Mínimos Quadrados Parciais, dividida em seis modelos (três sem restrição de comprimentos de onda e três em que os comprimentos de onda foram restritos a 1670-900 cm^{-1}). Nesta análise utilizámos as mesmas amostras que na Análise de Componentes Principais, e o método foi capaz de diferenciar as amostras negativas para a CoVid-19 das amostras positivas, com uma média de exatidão de 85%. Foi possível concluir que as zonas correspondentes ao comprimento de onda referentes a proteínas e carboidratos foram as principais características consideradas para classificar as amostras.

Os resultados sugerem que é possível distinguir entre amostras positivas e negativas, utilizando tanto a Análise de Componentes Principais como Análise Discriminante de Mínimos Quadrados Parciais, o que sustenta a hipótese de que a Espectroscopia de Infravermelho com Transformada de Fourier com Reflexão Total Atenuada pode ser uma metodologia promissora para o diagnóstico da doença de CoVid-19.

Professor Orientador: Doutor João Pedro Martins de Almeida Lopes, Professor Auxiliar

Palavras-chave: CoVid-19; Vacinas; Espectroscopia de infravermelho; Bio-fluídos; Plasma

Abstract

In this study, we evaluated the accuracy of outcomes, when using Fourier Transform Infrared with Attenuated Total Reflection spectroscopy, coupled with multivariate analysis, to analyze and identify plasma samples from patients with Coronavirus Disease 2019 positive PCR test and no reported positive PCR test.

Triage process of Coronavirus Disease 2019 has become a difficult process, since the increasing of cases and, consequently, the affluence to the hospitals. With this, the hypothesis of using this technic as a diagnostic tool could be a faster method to identify patients with higher inflammatory indicators.

Firstly, an exploratory analysis using Principal Component Analysis was performed utilizing the spectra of samples from diverse sources patients, vaccinated with Pfizer/BioNTech and AstraZeneca vaccines.

In a next phase, a Partial Least Squares-Discriminant Analysis was developed, divided into six models (three with no wavenumber restriction and three in which wavenumber were restricted to 1670-900 cm^{-1}). We employed the same samples as in Principal Component Analysis, and the method was able to differentiate the Coronavirus Disease 2019 negative samples from Coronavirus Disease 2019 positive samples, with an accuracy media of 85%. It was possible to conclude that proteins and carbohydrates wavelength correspondent zones were the main features considered to classify the samples.

The results suggest that it is possible to distinguish between positive and negative Coronavirus samples, using Principal Components Analysis and Partial Least Square-Discriminant Analysis, which supports the hypothesis that Transform Infrared with Attenuated Total Reflection spectroscopy can be a promising methodology for Coronavirus Disease 2019 diagnosis.

Advisor Professor: Doctor João Pedro Martins de Almeida Lopes, Assistant Professor

Keywords: CoVid-19; Vaccines; Infrared spectroscopy, Biofluids, Plasma

Acknowledgments

I would like to thank to everyone that contributed and supported me during the realization of this thesis.

First of all, I thank to my parents, my sister Sofia and my brother João, for understanding my busy and bad mood days and for allowing me to be 100% focused on this project.

To my advisor, Professor Dr. João Almeida Lopes, for giving me the opportunity to participate in subject Project I, where I started to learn everything I needed to accomplish this thesis. I would like to thank all the knowledge and support, for always giving me the best conditions to work and for understanding my availability.

To all members of the Molecular Microbiology and Biotechnology group at iMed.ULisboa, especially Professor Dr. João Gonçalves, Carlos Araújo and Igor Casado. To Joana, Nuno and Inês, from the laboratory, for always solving everything I needed and supporting me. To Professor Dr. Helena Florindo and her team, for allowing me to pick up my samples every hour I needed, even if during lunch time or after 8 p.m..

I would like to thank Nádia Aguiam for all the guidance and life advices.

To all my friends who accompany me since the beginning of this journey, Ana Filipa, Beatriz, Joana, Filipe, Lina, Mariana B., Mariana M., Patrícia and Vânia.

Final but not least, I would like specially to thank my oldest friends, that were always by my side and always made me believe that I could handle everything, to Tiago, Catarina and Cláudia.

Abbreviations

ATR-FTIR - Attenuated Total Reflection Fourier Transform Infrared

BD - Becton & Dickinson

B's - Biological Replicates

CoVid-19 - Coronavirus Disease 2019

CV - Cross-validation

EDTA - ethylenediaminetetraacetic acid

FIR - Far Infrared

ICU - Intensive Care Unit

I's - Instrumental Replicates

IR - Infrared

IRE - Internal Reflection Element

MIR - Mid Infrared

NIR - Near Infrared

NK - Natural Killer cells

PCA - Principal Component Analysis

PCs - Principal Components

PLS - Partial Least Square

PLS-DA - Partial Least Square-Discriminant Analysis

RT-PCR - Reverse-Transcriptase Polymerase Chain Reaction

SARS-CoV-2 - Severe Acute Respiratory Syndrome Coronavirus 2

VIP - Variable Importance in Projection

Index:

Figure Index:

Table Index:

1 Introduction

Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy is a non-destructive technique that can be applied to a vast range of applications (1), from food chemistry (2) to biological samples (3) or industry (4). In the vibrational spectroscopy field, there are a lot of studies that demonstrate the potential of these class of techniques as a clinical diagnostic tool, using biofluids vibrational properties (5,6). It has been demonstrated that serum and plasma appear as an ideal intermediate for routine clinical use, once they are easily accessible and collectable by an almost non-invasive (7) and low-cost (8) method. In addition, it is well known that alterations of plasma proteins are good indicators of physiopathological changes, caused by a large range of diseases, including viral infections such as coronavirus disease 2019 (CoVid-19) (9).

1.1 CoVid-19

Healthcare providers typically stratify CoVid-19 patients based on clinical presentations, such as symptoms, peripheral pulse oxygen saturation, and blood pressure (10), dividing patients, according to disease progression, into two groups: asymptomatic or mild cases that usually recover and severe cases that develop multi organ and respiratory failure, requiring intensive care unit (ICU) admission (11,12) . A retrospective observational study concluded that, in 2020, common clinical features of patients with CoVid-19 included fever (83%), cough (82%), shortness of breath (31%), and muscle ache (11%) (13).

Clinical assessment is indispensable but sometimes subjective. On the other hand, laboratory markers can provide additional, objective information which can significantly impact many components of patient care quality. These have various potential benefits, such as identification and confirmation of at-risk patients with stratification of CoVid-19 severity and assistance in the establishment of care criteria (14). For example, some studies suggest that eosinopenia along with lymphopenia may be a useful indicator for diagnosing CoVid-19 in those patients with typical symptoms and radiological changes (13,15), along with a significant relationship between the disease severity and the levels of proinflammatory cytokines and subsets of immune cells, with a severe reduction in the frequency of CD4+ and CD8+ T cells, B cells and natural killer (NK) cells (16–18). Dorgham *et al* showed that critically severe patients do not have higher viral load than

less severe patients, but rather exhibit higher levels of inflammatory cytokines and lower type-I interferon response (19).

Due to the wide range of symptoms that overlap with other respiratory infections, diagnosis usually depends on laboratory detection by reverse-transcriptase polymerase chain reaction (RT-PCR) to identify ongoing infection, despite not being the fastest technique. Beside diagnosis, it is particularly important to monitor the neutralizing capacity of the antibodies, which is important to assess the protection expected against an infection upon re-exposure to the virus (20).

1.2 ATR-FTIR spectroscopy

ATR-FTIR is a vibrational spectroscopic technique, where IR light is directed through an internal reflection element (IRE) with a high refractive index (for example diamond) and penetrates approximately 0.5-5 μm in the sample, counting from the surface of the IRE (**Figure 1**). This can be used for the analysis of biofluids, which must be in intimate contact with the IRE surface, due to biomolecules exhibit different responses to a vast range of wavenumbers of light, which allows for an objective identification and quantification of compounds depending on their molecular composition (1,21,22).

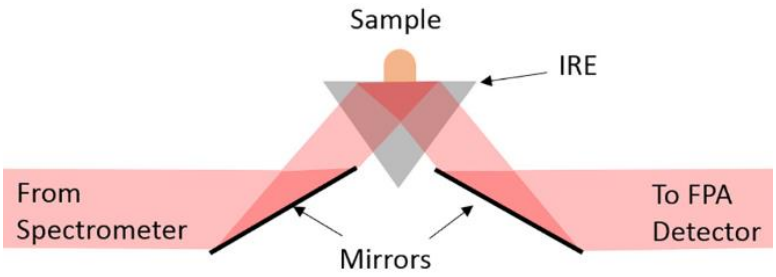


Figure 1- Diagram of the ATR-FTIR sampling measurement principle. Adapted from (1)

Infrared (IR) region (10000-100 cm^{-1}) can be divided in Near Infrared (NIR), Mid Infrared (MIR) and Far Infrared (FIR). MIR is the most used when analyzing biological samples and includes wavenumbers between 4000 and 400 cm^{-1}. The studied spectral range of 1670-900 cm^{-1} is included in the region called the "bio fingerprint region" (1800-900 cm^{-1}), which offers the most information on the chemical compounds of biological samples, including amide I, II and III proteins (1700-1500 cm^{-1}, 1350-1200 cm^{-1}). The amide I band (between 1600 and 1700 cm^{-1}) is mainly associated with the C=O stretching

vibration (70-85%) and is directly related to the backbone conformation. Amide II results from the N-H bending vibration (40-60%) and from the C-N stretching vibration (18-40%). Other relevant wavenumber regions are those related with nucleic acids (1200-1000 cm-1), carbohydrates (1200-900 cm^{-1}) and phosphodiester stretching bands (901 cm^{-1}, 1506 cm^{-1}) (23,24).

The potential of spectroscopic techniques for the detection and identification of virus-infected cells, and the ability to discriminate between contaminated and non-contaminated cells, has been studied using statistical methods as a sensitive, rapid and reliable methodology (25). ATR-FTIR spectroscopy has been proved to be capable of detecting metabolic changes after viral infection, which can be justified by the increase of immunoglobulin levels along with an immune dysregulation and high level of proinflammatory cytokines, to fight the infection, including response to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (24,26). Collection of an FTIR absorption spectrum involves collecting a single-time spectrum. On the other hand, a spectral map is composed of many points spectra acquired step by step. To account for the variation of atmospheric conditions over the extended acquisition time, it is necessary to set up background scans to be taken at defined intervals, or between each sample, to reduce the effects of these constant changes (27).

1.3 Chemometrics

1.3.1 Principal Component Analysis (PCA)

One of the most utilized methods for spectral data analysis, when exploratory data analysis is the goal, is the Principal Component Analysis (PCA) method (28). This method decomposes the spectra and then reconstruct them from only a few of their principal components (PCs), leading to a simpler representation. PCs describe the major sources of variance among a spectral data set and are ordered by the amount of variance they explain, which means that the first PC (PC1) describes the majority of the data. The choice of an optimal number of PCs, balances the necessity to explain as much of the original data as possible, with the need to avoid including too much noise into the new representation (overfitting). The outcomes are the scores (how much each PC is described in samples) and loadings (the PCs) (28–30).

1.3.2 Partial Least Square Discriminant Analysis (PLS-DA)

Unlike PCA, PLS-DA considers not only an independent set of descriptors (X) but also the dependent variables (Y), into which data are projected, being categorized as a discriminant method that produces predicted values for each of the Y-variables. Each row in Y is a vector encoding class membership information (28,30). To ensure a correct number of components (PCs) to describe the data, it is crucial to validate the model. While picking an insufficient number of components can incur in the risk of not explaining all the pertinent variance (underfitting), including too many of them (probably containing noise), can lead to overfitting. Cross-validation is the most utilized method to prevent underfitting and overfitting, and it is particularly appropriate when there is a small number of available samples. This consists in leaving out some of the data while building a model, and then validate it applying the model to the removed data (28,30).

1.4 SARS-CoV-2 vaccines

Among all the control measures that aim to prevent viruses spread, such as washing hands, social distance or face masks, the implementation of vaccines against SARS-CoV-2 is a main strength in slowing down the coronavirus disease 2019 (CoVid-19) pandemic. Between those who were included in IV phase clinical trials (real-world), the most utilized were Pfizer/BioNTech (BNT16b2), AstraZeneca (AZD1222 ChAdOx1 nCoV-19), Moderna (mRNA-1273) and Johnson&Johnson (Ad26.COV2.S) vaccines, whose characteristics are resumed in **Table 1.** Pfizer/BioNTech and Moderna vaccines utilizes a mRNA platform encrypting the spike protein of SARS-CoV-2 (structure responsible for the recognition of the virus) in order to stimulate the immune system. On the other hand, AstraZeneca and Johnson&Johnson vaccines take advantage of a viral vector (adenovirus encoding the spike SARS-CoV-2 glycoprotein) that are sufficient to induce host immune responses but cannot replicate inside host cells since they are blocked for DNA synthesis (31,32). According to diverse studies about safety and efficacy monitoring, reported serious adverse events were very uncommon and the relation benefits-risks of COVID-19 vaccination is truly positive (31,33).

Table 1 – Most utilized vaccines resume. Adapted from (31)

	Manufacturer	**Type**	**Base Composition**
BNT16b2	Pfizer/BioNTech	RNA-based	Synthetic messenger ribonucleic acid (mRNA) encoding the spike protein of SARS-CoV-2.
AZD1222 ChAdOx1 nCoV-19	AstraZeneca/ University of Oxford	Non-replicating viral vector	Chimpanzee Adenovirus encoding the SARS-CoV-2 spike glycoprotein.
mRNA-1273	Moderna	RNA-based	Synthetic messenger ribonucleic acid (mRNA) encoding the spike protein of SARS-CoV-2.
Ad26.COV2.S	Johnson&Johnson	Non-replicating viral vector	Replication-incompetent recombinant adenovirus vector expressing the SARS-CoV-2 spike protein.

2 Materials and methods

2.1 Plasma collection

In this study, we utilized a total of 60 plasma samples, obtained from patients admitted in diverse social and healthcare institutions, divided into those who reported no positive PCR test for CoVid-19 ("negative" samples) and those with positive PCR test for CoVid-19 ("positive" samples). Blood was collected by venipuncture to BD (Becton & Dickinson) vacutainer® blood K2 ethylenediaminetetraacetic acid (EDTA) collection tubes. To isolate plasma, blood was centrifuged at 2000g for 10 minutes at room temperature. Plasma samples were aliquoted and stored at -20ºC until use.

2.2 Sample preparation and spectral collection

Plasma samples were removed from storage and centrifuged for 3 minutes at 2000g. After centrifugation, 10µL was pipetted and performed as resumed in **Table 1**. Infrared spectra were collected using the Thermo Scientific™ Nicolet™ IS5 FTIR Spectrometer, with a diamond crystal Thermo Scientific™ iD5 ATR Accessory (**Annex 1**), in the range of 4000-450 cm^{-1}, at a resolution of 2 cm^{-1}, with 16 co-added scans **(Figure 1).**

A background spectrum was performed before each new sample scan, which was automatically subtracted from the acquired one. Samples were pipetted directly to the ATR crystal (after its cleaning with isopropanol) and, to minimize the water contribution to the spectrum, the sample was left to dry at room temperature on the crystal's surface for 30 minutes. Spectra were collected every 5 minutes, resulting in six instrumental replicate acquired for each biological replicate (I's), with the objective to confirm that the final signal is representative of the sample. Spectra are identified according to the following structure: ID Sample-Biological Sequence (B)-Instrumental Sequence (I)-Drying Time (T).

Table 2 - Method description resume

	Drying time	**Drying temperature**	**Scanning Time**	**Scanning Types (each sample)**
Directly to the ATR crystal	30 minutes	Room Temperature	Every 5 minutes	1 Biological (B's) 6 Instrumental (I's)

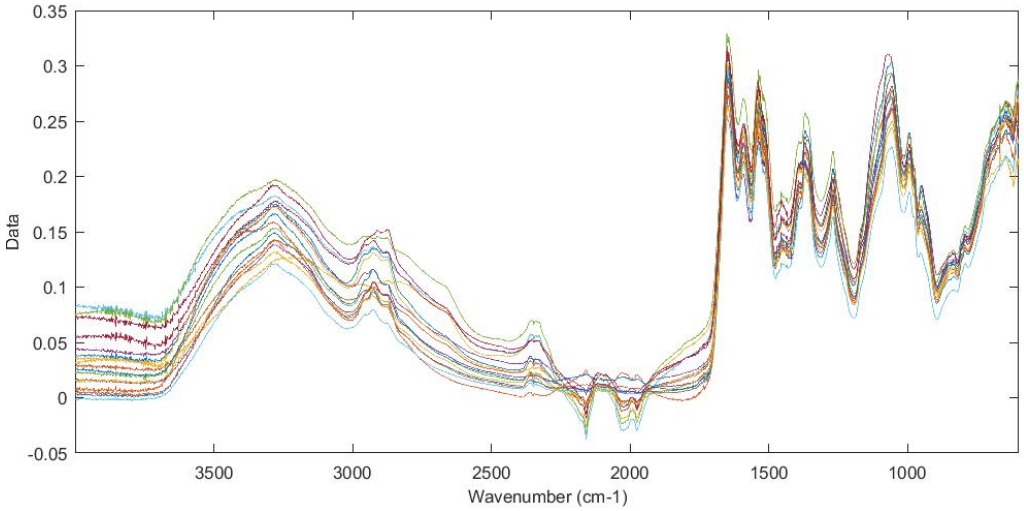


Figure 2 - Representation of all collected ATR-FTIR spectra (collected at the time point 30 minutes).

2.3 Spectral data analysis

Data were pre-processed and analyzed using PCA and PLSDA. The multivariate exploratory analysis method PCA was adopted to assess the degree of correlation between the variables and to allow a better and simpler visualization of the intrinsic relationships within this dataset. PLSDA allows us to classify and categorize samples, based on a model group. The processed IR spectra were restricted to the biologically relevant region (1670-900 cm^{-1}) and the data were treated using MATLAB R2016b and Partial Least Square (PLS) Toolbox (Eigenvector Research Inc. ®). The spectra were pre-processed using the first derivative (Savitzky-Golay method), in order to baseline offset variations (Savitzky-Golay method with 15 time points, 2nd order polynomial and first derivative). Before application of PCA and PLSDA, all spectra were mean-centered. Hotelling's T2 and squared residuals statistics were analyzed in order to identify any abnormal spectral measurement.

3 Results and Discussion

3.1 PCR and Nucleocapsid results

Initially, 60 plasma samples were equally divided into two groups, according to the type of vaccine administered (AstraZeneca and Pfizer-BioNTech vaccines). With the objective to confirm and understand PCR CoVid-19 results, a Nucleocapsid study was performed, by the Molecular Microbiology and Biotechnology group at iMed.ULisboa, and a threshold value was defined to categorize samples in positive or negative for CoVid-19 (see **Annex 2**). In **Table 2** we can observe that, despite the study confirms almost every PCR test information, there are two samples where PCR and Nucleocapsid results are not in agreement, (VC1361 and VC1366). These samples will be highlighted in grey color in the following charts.

Table 3 - Samples resume with PCR and Nucleocapsid results for CoVid-19.

Pfizer-BioNTech	PCR	Threshold	AstraZeneca	PCR	Threshold
VC1,4			VC1335		
VC2,4	+	+	VC1336	+	+
VC3,3			VC1337	+	+
VC4,3			VC1338	+	+
VC5,3			VC1339	+	+
VC6,3			VC1340	+	+
VC7,3			VC1341		
VC8,3			VC1342	+	+
VC9,3			VC1344	+	+
VC10,3			VC1345		
VC11,3			VC1346		
VC12,3	+	+	VC1347	+	+
VC13,3			VC1349		
VC14,3	+	+	VC1351		
VC15,3			VC1352		
VC16,3	+	+	VC1353	+	+
VC17,3			VC1354	+	+
VC18,3			VC1355	+	+
VC19,3			VC1356	+	+
VC20,3			VC1357	+	+
VC21,3			VC1358		
VC22,3			VC1359	+	+
VC26,3			VC1360		
VC27,3			VC1361		+
VC28,3	+	+	VC1362		
VC45,3	+	+	VC1363	+	+
VC51,3	+	+	VC1365		
VC74,3			VC1366		+

VC98,3	+	+	VC1367		
VC108,3	+	+	VC1368		

When analyzed the distribution of samples according to Nucleocapsid study values and defined threshold (orange line in **Annex 2**), it was noted that some samples appear near to limit value, which can lead to misleading results due to the proximity of the values.

3.2 Full spectra PCA

It is well stablished that biomolecules exhibit different vibration responses when submitted under a vast range of wavenumbers of light. With this, we can identify the main composition of a sample, by analyzing its IR spectrum. As we can see in **Figure 4**, it is possible to identify protein related zones, such as N-H stretching around 3300 cm^{-1}, amide I band at 1700 cm^{-1}, Amide II band at 1600 cm^{-1} and C=O stretching at 1400 cm^{-1} (protein side chains). Another band appears around 1200-900 cm^{-1}, referring to the C-O-C ring vibrations of carbohydrates. It is also possible to observe the C-H vibrations of fatty acids (from lipids) between 2800 and 3000 cm^{-1}, and the PO^{2-} stretching bands (phosphodiester bounds from DNA/RNA or phospholipids) around 1300 cm^{-1}.

Pre-processed PC1 loading with first derivative is presented in **Figure 3**, where we can observe that the samples variations are more evident in the region between 1700-1400 cm^{-1}, indicating that the proteins composition is the main indicator of variability.

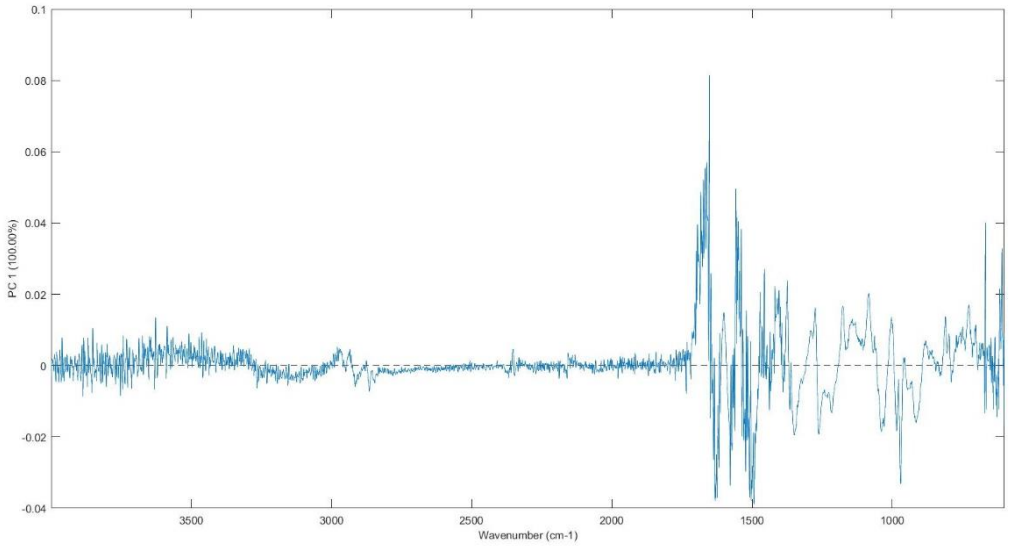


Figure 3 - ATR-FTIR PC1 loading of plasma sample VC1365-B1-I6-T30 after pre-processing (first derivative).

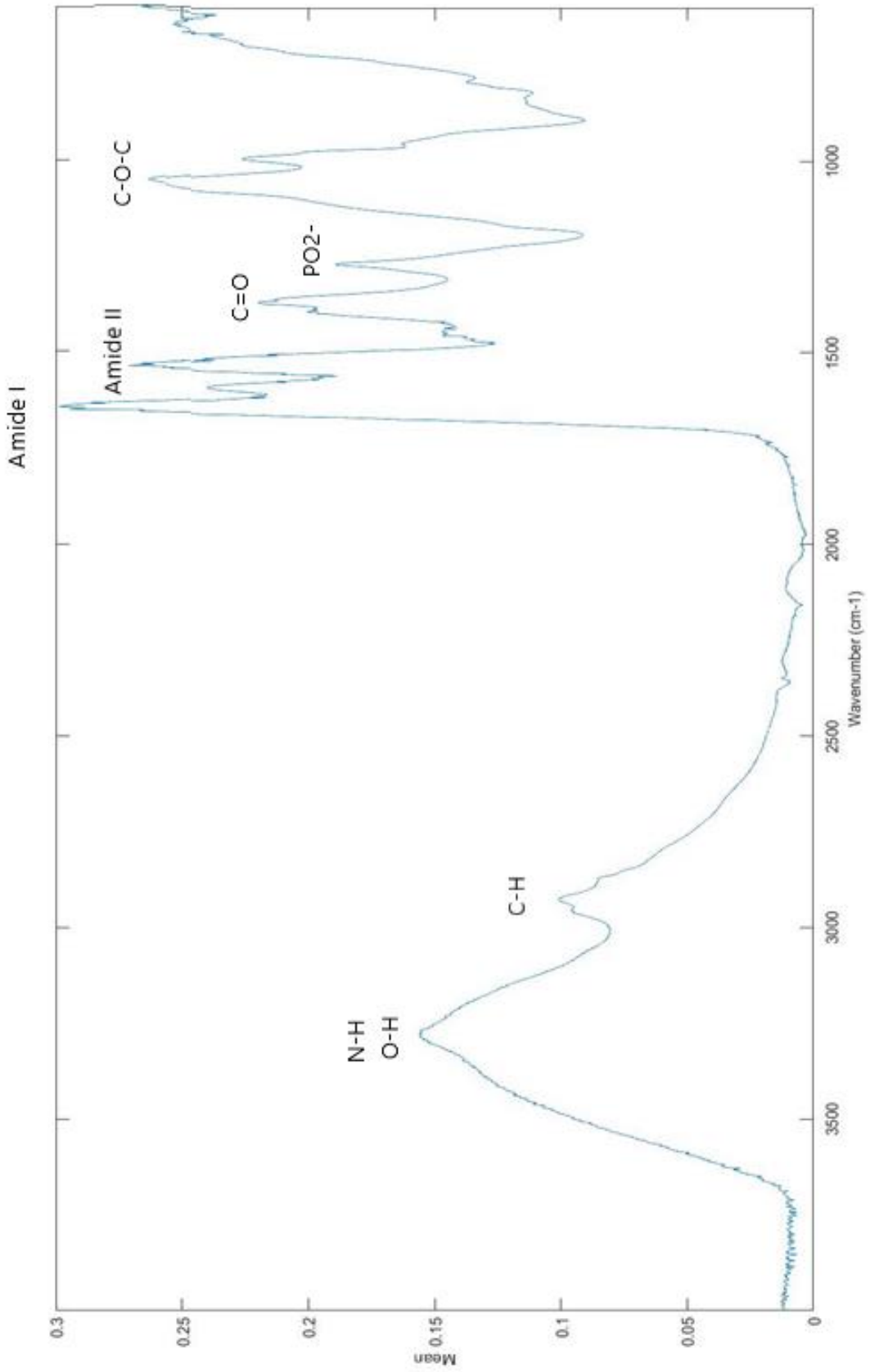


Figure 4 - Wavenumber zones of IR spectrum VC1365-B1-I6-T30 related to molecules identification.

In a first step, all spectra were analyzed by PCA, with no previous identification, categorization, or wavenumber restriction. There were two samples (VC74.3 and VC1347) that were negatively influencing the model (see **Annex 3** and **Annex 4).** With the objective of preventing misleading results, these samples were excluded for the remaining analysis steps. In **Figure 5** we can see first phase analysis when excluding samples VC74.3 and VC1347.

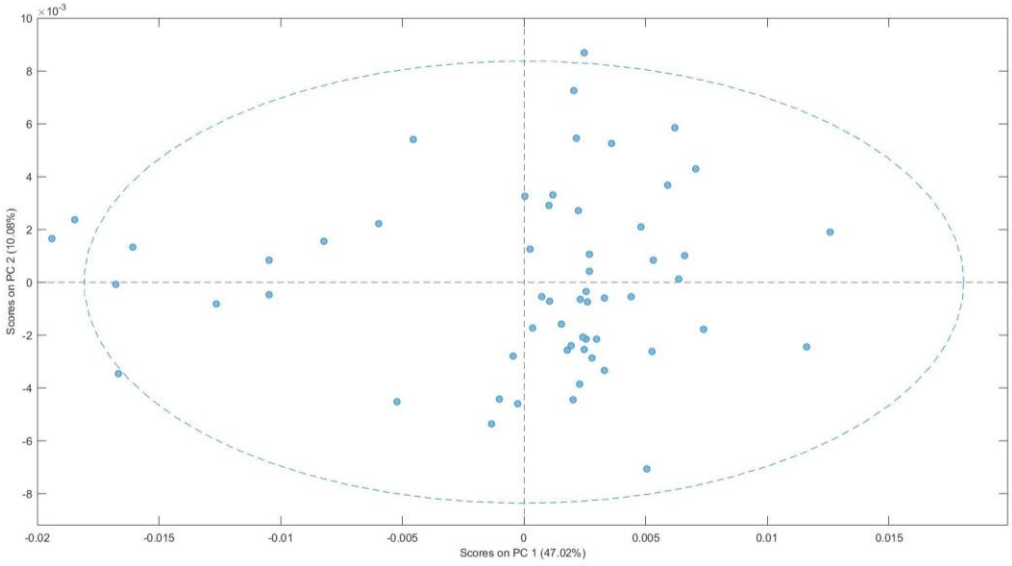


Figure 5 – Scores plot (PC1, PC2) of the PCA model with no wavenumber restriction.

3.3 Wavenumber restriction PCA

Once the biologically relevant region to analyze plasma samples is the "bio fingerprint region", we performed the next series of analyses with restricted wavenumbers between 1670 cm^{-1} and 900 cm^{-1}. In **Figure 6**, we can distinguish two groups of samples, where the samples VC22.3, VC27.3, VC28.3, VC98.3 and VC108.3 are slightly outside the confidence limit.

Besides viral infections, ATR-FTIR is also utilized as disease diagnostic tool once all metabolic changes reflect into plasma composition. Among these outliers, it was observable that samples VC27.3 and VC28.3 presented a yellow color, comparing to the rest almost transparent samples. It can suggest some dysregulation that may justify the result. About the other samples, it is not possible, with the information available, to indicate a reason for the observation.

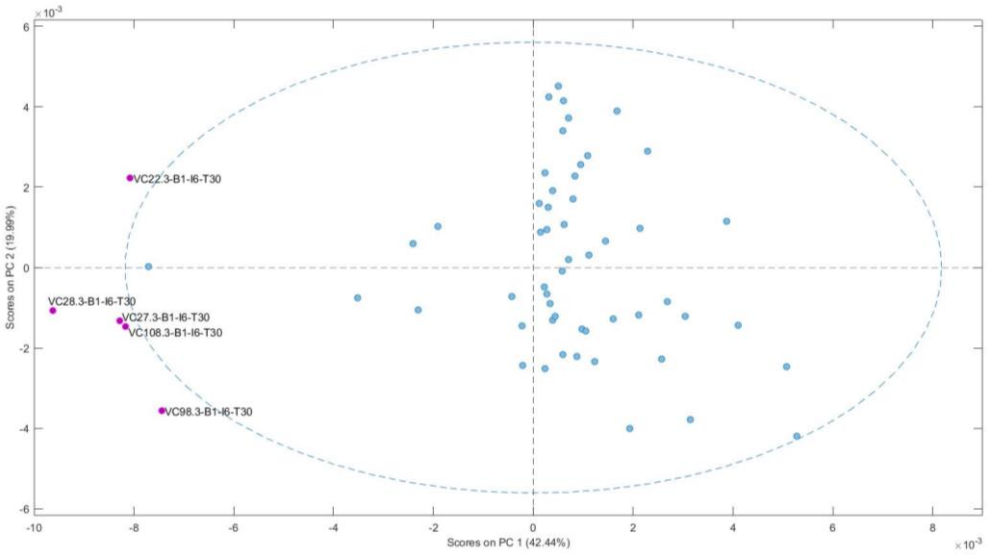


Figure 6 - Scores plot (PC1 and PC2) of the PCA model with wavenumber restriction to 1670-900 cm^{-1}.

3.4 CoVid-19 Positive versus Negative samples

With the objective to investigate whether there is information in scores able to separate positive and negative CoVid-19 samples, they were categorized according to the respective result (red for CoVid-19 positive samples and green for CoVid-19 negative one). In **Figure 7**, it is quite defined that samples are being separated according to PC2 (see **Annex 5**), once we can observe a red area for positive PC2 scores, and a green area for negative PC2 scores.

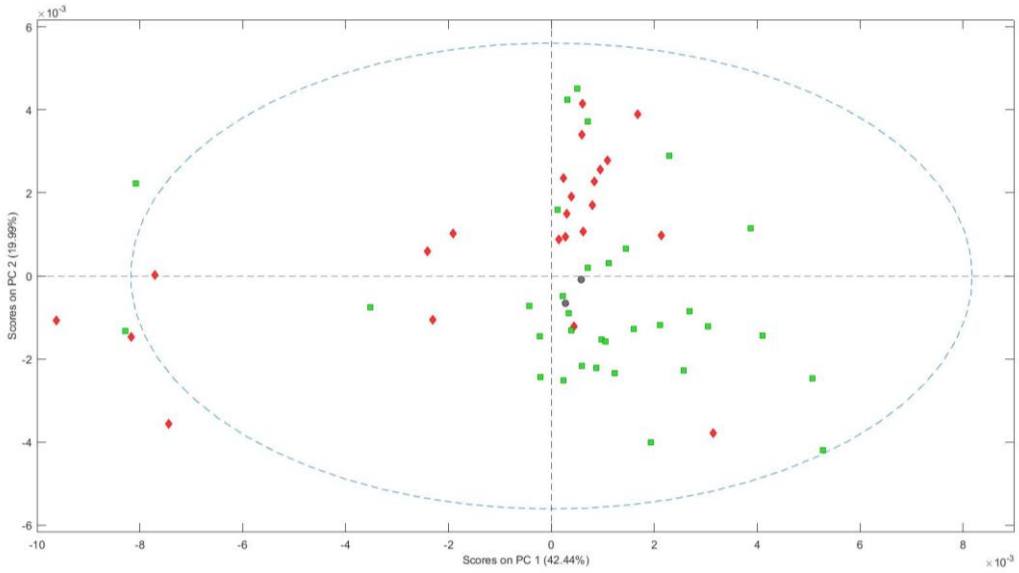


Figure 7 - PCA scores plot (PC1 and PC2) showing wavenumber restriction to 1670-900 cm^{-1} in positive (red) and negative (green) samples.

With this, a loading PC2 analysis (**Figure 8**) was performed to understand what wavenumber region is mostly responsible for the separation observed. Here we can conclude that wavenumbers between 1670-1300 cm^{-1} are the main influence, which refers mostly to proteins associated areas. A similar result, but less intense (mostly noise) was obtained with a Hotelling's T2 statistical analysis, presented in **Figure 9**, which allows us to identify the spectral regions that justify the separation between two groups of samples (reference group (red) and test group (purple) represented in **Annex 6).**

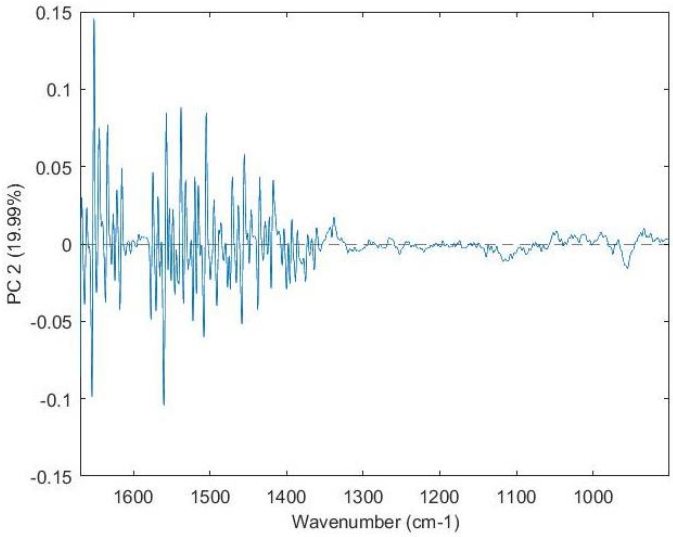


Figure 8 – ATR-FTIR PC2 loading, showing the 1700-1300 cm^{-1} (proteins) influence.

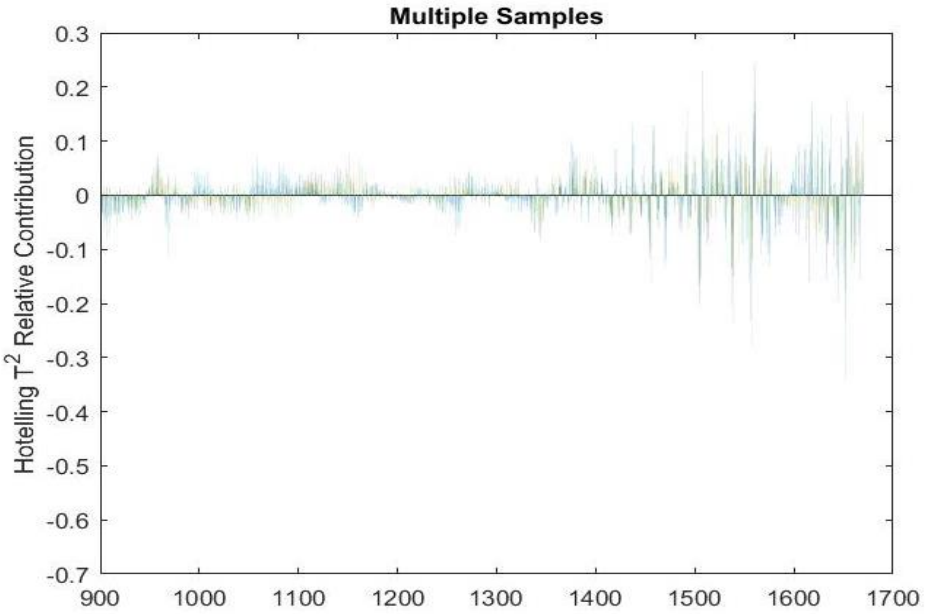


Figure 9 – Hotelling's T2 statistic, showing the 1700-1400 cm^{-1} (proteins) influence.

Of the samples that appear as outliers, according to their CoVid-19 result (see **Annex 7**), samples VC15.3, VC22.3, VC28.3, VC98.3, VC108.3 and VC1353 have their nucleocapsid study results near to threshold, which can indicate some possibility of false positive or false negative tests. It can also happen because of no exacerbation of immunological system, or in case of some metabolic dysregulation. The fact that are being considered samples from patients vaccinated with both vaccines in study also creates variability.

3.5 CoVid-19 positive samples

Comparing all CoVid-19 positive samples, in **Figure 10** we can see that they tend to converge to the same area, showing a very defined red blot. Despite that, there is a group of samples and an isolated one (VC2.4) that appear separated from the others. Except sample VC45.3, on which there is not a direct explanation, the rest of the samples are the same that appeared previously as outliers.

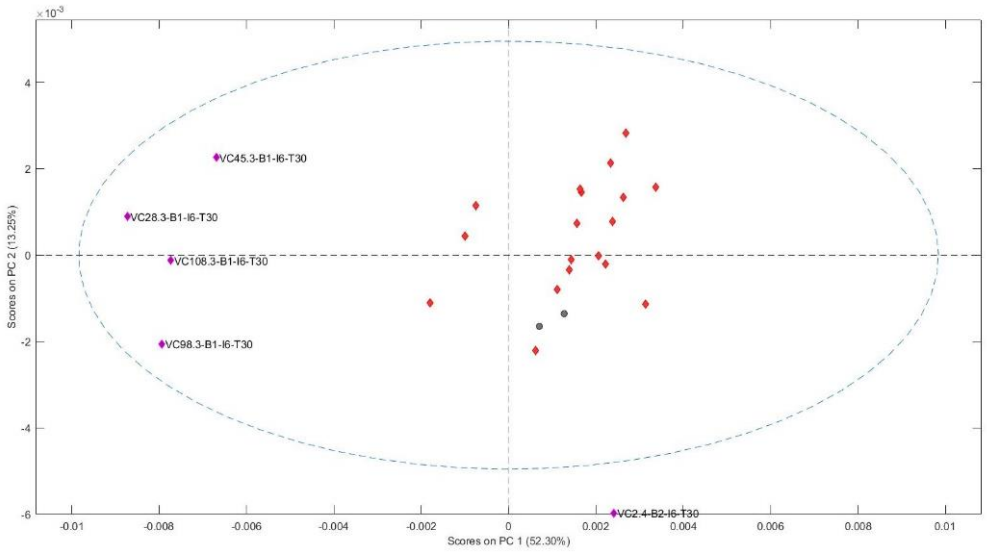


Figure 10 - PCA scores plot (PC1 and PC2) with wavenumber restriction to 1670-900 cm^{-1} in CoVid-19 positive samples.

Figure 11 shows the contribution plot for the Hotelling's T2 statistic. Contributions highlight differences between groups of samples and allow to investigate which wavenumbers are the most responsible for some observed difference. The left chart indicates that, for the group of samples on the left (Fig. 10), comparing to the red blot, most of the discrimination is justified by the wavenumbers 1100-1000 cm^{-1} and 1200-

1100 cm^{-1} (mostly carbohydrates). The right side chart of Fig. 11 refers to sample VC2.4, and also shows a 1200-1100 cm^{-1} statistical importance, adding a 1400-1300 cm^{-1} (C=O from carboxylate in protein side chains) and some 1650-1500 cm^{-1} (Amide I and Amide II from proteins) influence.

a. b.

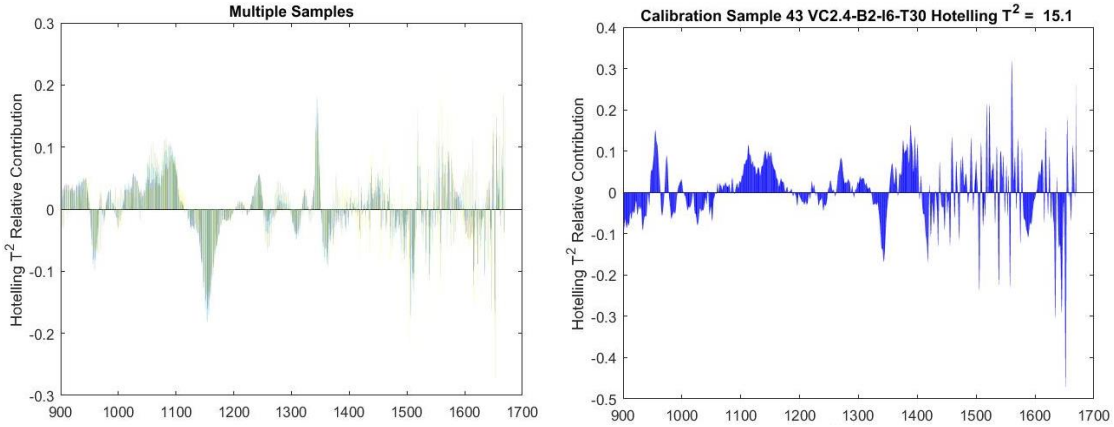


Figure 11 – Hotelling's T-Square statistical analysis comparing the red blot with samples on the left of Fig. 10 and with sample VC2.4. (Figure 11-a and Figure 11-b, respectively).

3.6 CoVid-19 negative samples

The analysis of plasma samples from patients for which no positive CoVid-19 test was known (the "negative" samples) show a pattern apparently formed by randomly distributed samples around the model's center (**Figure 12).**

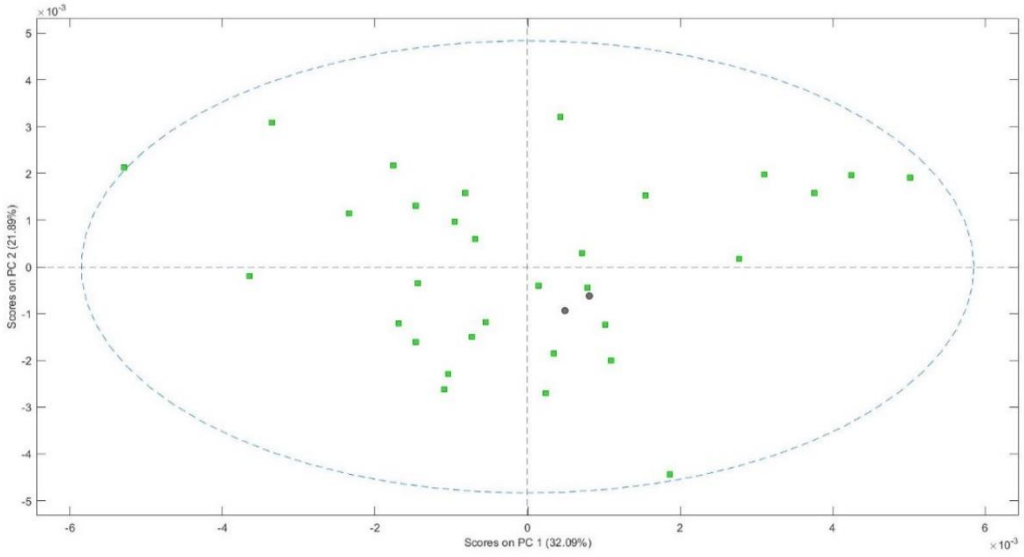


Figure 12 - PCA scores plot (PC1 and PC2) with wavenumber restriction to 1670-900 cm^{-1} in CoVid-19 negative samples.

The contributions to the Hotelling's T2 statistic is presented in **Figure 13**, with three different groups of samples analyzed against a reference group (**Fig 13-a**). As we can observe in **Fig. 13-b.**, the sample group that appears on the left of the reference group is predominantly being influence by wavenumbers between 1200-1000 cm^{-1} (carbohydrates). **Fig 13-c** shows that, in sample VC1341, is being taking into account mostly region between 1250-900 cm^{-1} (carbohydrates) with some 1650-1500 cm^{-1} (Amide I and Amide II from proteins) influence. Analyzing **Fig 13-d**, samples placed on the right are mainly being influenced by noise signals.

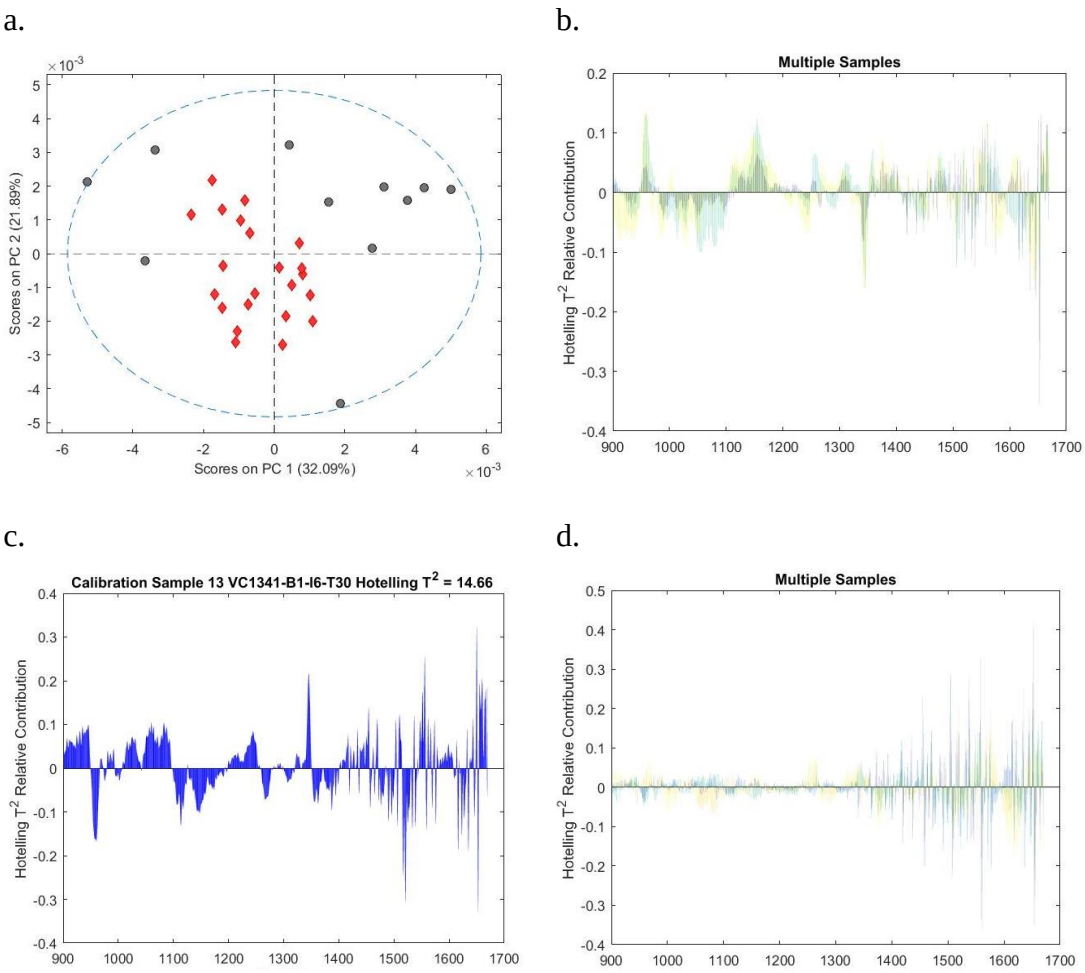


Figure 13 – Hotelling's T-Square statistical analysis comparing a reference group, represented with a red blot in Fig 13-a, with a sample group on the left of reference group, a sample VC1341 and with a **s**ample group on the right of reference group (Fig. 13-b, Fig. 13-c and Fig. 13-d, respectively).

3.7 Pfizer-BioNTech vaccinated individuals analysis

As said previously, CoVid-19 positive and negative samples were grouped into those vaccinated with Pfizer-BioNTech and those with AstraZeneca, to compare them. The first group is presented in **Figure 14**, where it is possible to distinguish two defined blots. On the left group predominates positive samples (red) and on the right group, prevails negative samples (green).

As we can see in **Figure 14**, some samples appear as outliers, according to their CoVid-19 result (see **Annex 8)**. Among positive outliers (positive samples that behave as negative samples) only sample VC16.3 has a viable explanation, that is the fact that its nucleocapsid study value appear close to threshold. Evaluating negative outliers (qualities similar to positive samples), sample VC27.3 is the yellow sample presented earlier and VC22.3 also has its nucleocapsid value near to threshold. Statistically analyzing, from the 29 original spectra, the model grouped correctly 22, so we have a 75.9% of accuracy

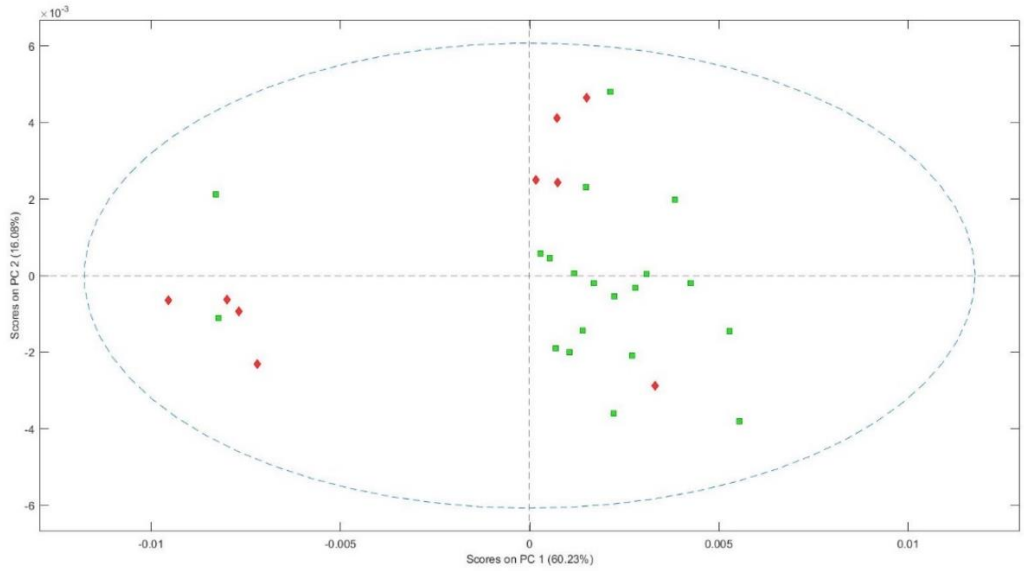


Figure 14 - PCA scores plot (PC1 and PC2) with wavenumber restriction to 1670-900 cm-1 in samples from patients vaccinated with Pfizer-BioNTech vaccine.

A Hotelling's T-Square statistical analysis is presented in **Figure 15**, with two different groups (sample groups on the left and above the red blot identified in **Annex 9**) analyzed against a reference group (red blot in **Annex 9**). **Fig. 15-a** indicates that, in the left group, is being considered mostly region between 1200-900 cm^{-1} (carbohydrates) with some 1500-1350 1200-900 cm^{-1} (Amide I and Amide II from proteins) influence. On the other hand, in **Fig. 15-b** we observe only a 1200-900 cm^{-1} less intense effect, with mainly noise associated.

a. b.

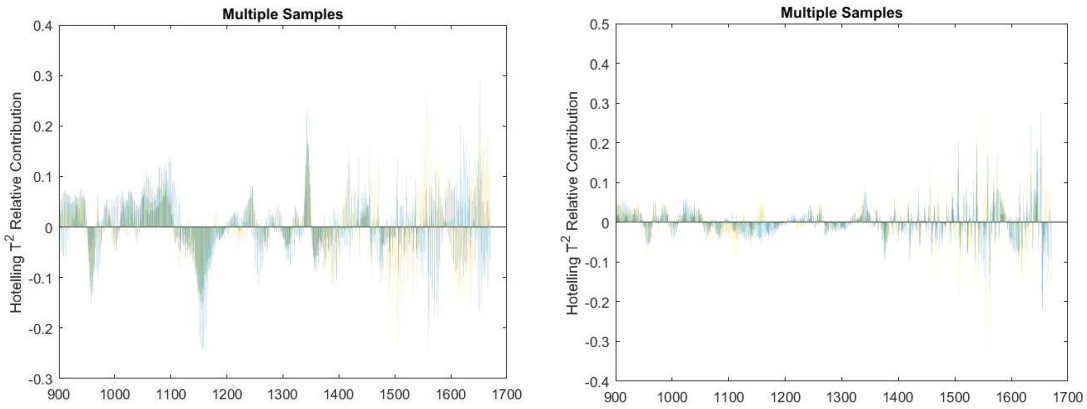


Figure 15 - Hotelling's T-Square statistical analysis comparing a reference group represented with a red bolt in **Annex 9** with a sample group on the left of reference group and a sample group above the reference group (Fig. 15-a and Fig. 15-b, respectively).

3.8 AstraZeneca vaccinated individuals analysis

Patients vaccinated with AstraZeneca vaccine are shown in **Figure 16**, where it is possible to distinguish one defined blot, composed by negative samples and one positive outlier, and some positive samples with negative outliers spread inside the model confidence limits. The propensity to positive samples be more distributed may indicate different degrees of infection and, consequently, distinct inflammatory system activation. Outliers' identification is found in **Annex 10**, but there is no apparent reason that justifies the observed outcomes.

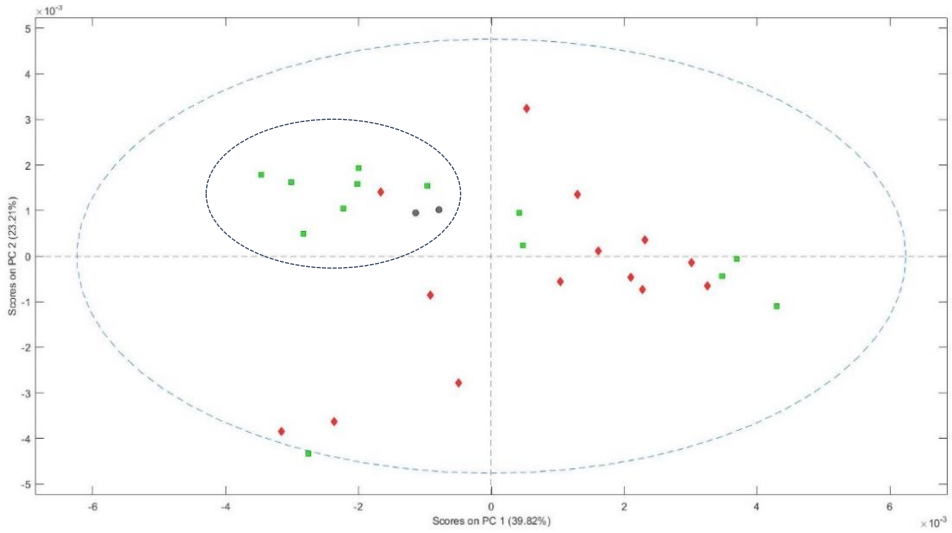


Figure 16 - PCA scores plot (PC1 and PC2) with wavenumber restriction to 1670-900 cm^{-1} in samples from patients vaccinated with AstraZeneca vaccine.

In order to understand the results, a Hotelling's T-Square statistical analysis is presented in **Figure 17**, with two different groups (sample groups beneath and on the left of the red blot identified in **Annex 11**) analyzed against a reference group (red blot in **Annex 11**). Here it is possible to conclude that the samples position is similar to that find in previous analysis. **Fig. 17-a** indicates that, in samples beneath the reference group, is being taking into account region between 1300-1100 cm^{-1} (carbohydrates) with some 1580-1500 cm^{-1} (C=O from proteins) influence. **Fig 17-b** does not show any significant result that can be interpreted.

a. b.

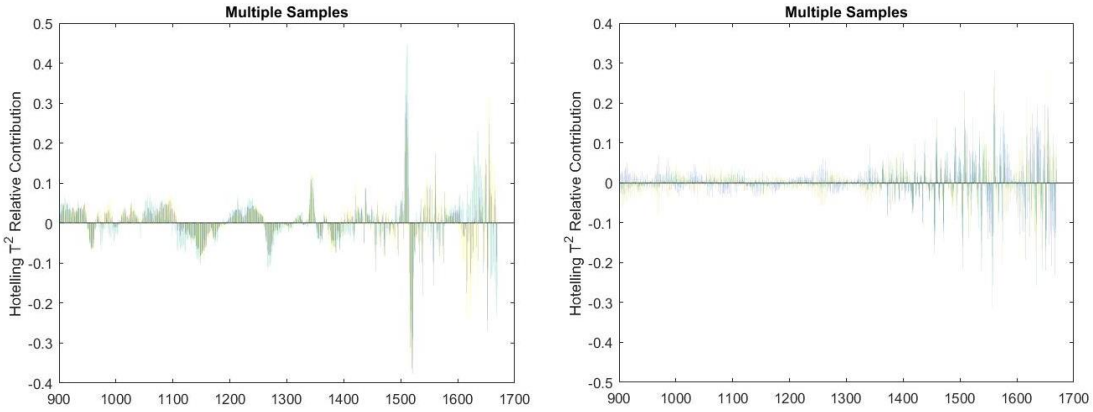


Figure 17 - Hotelling's T-Square statistical analysis comparing a reference group represented with a red blot in **Annex 11** with a sample group beneath the reference group a sample group on the left of the reference group (Fig. 17-a and Fig.17-b, respectively).

3.9 Partial Least Square Discriminant Analysis

PLS-DA is classified as a discriminant method that permits a categorization of samples. Three analyses were performed with three distinct model and tests groups, with the objective of validating the results. Each assay was divided into no restriction wavenumber models (models A, B and C) and 1670-900 cm^{-1} wavenumber restriction models (models A1, B1 and C1), resumed in **Annex 12**.

10 spectra from the original 56 (since the 2 spectra highlighted with grey color were not included), chosen almost randomly, were excluded from each model concept, and then used as validation (**Annex 12**). For the interpretation of each model, Strict Class Prediction (shows the predicted class for every single test sample) and Class Prediction Probability (presents the probability of one test sample to belong the predicted class) were

selected. Aiming to realize which wavenumbers areas are being considered in each model, Variable Importance in Projection (VIP) scores Y were performed.

Figure 18 presents model A results (**Fig.18-a** and **Fig 18-b**) and A1 results (**Fig.18-c** and **Fig 18-d**). As it is possible to see, there are no major differences between use no wavenumber restricted spectra or restricted to 1670-900 cm^{-1}. The principal variance is in the Class Prediction Probability, since in **Fig 18-d** there are more spectra near probability 50%, which indicates less certainty in outcomes of samples 5, 6 and 9 (VC12.3, VC13.3 and VC1337 respectively). Statistically talking, accuracy is 90% and 80%, respectively.

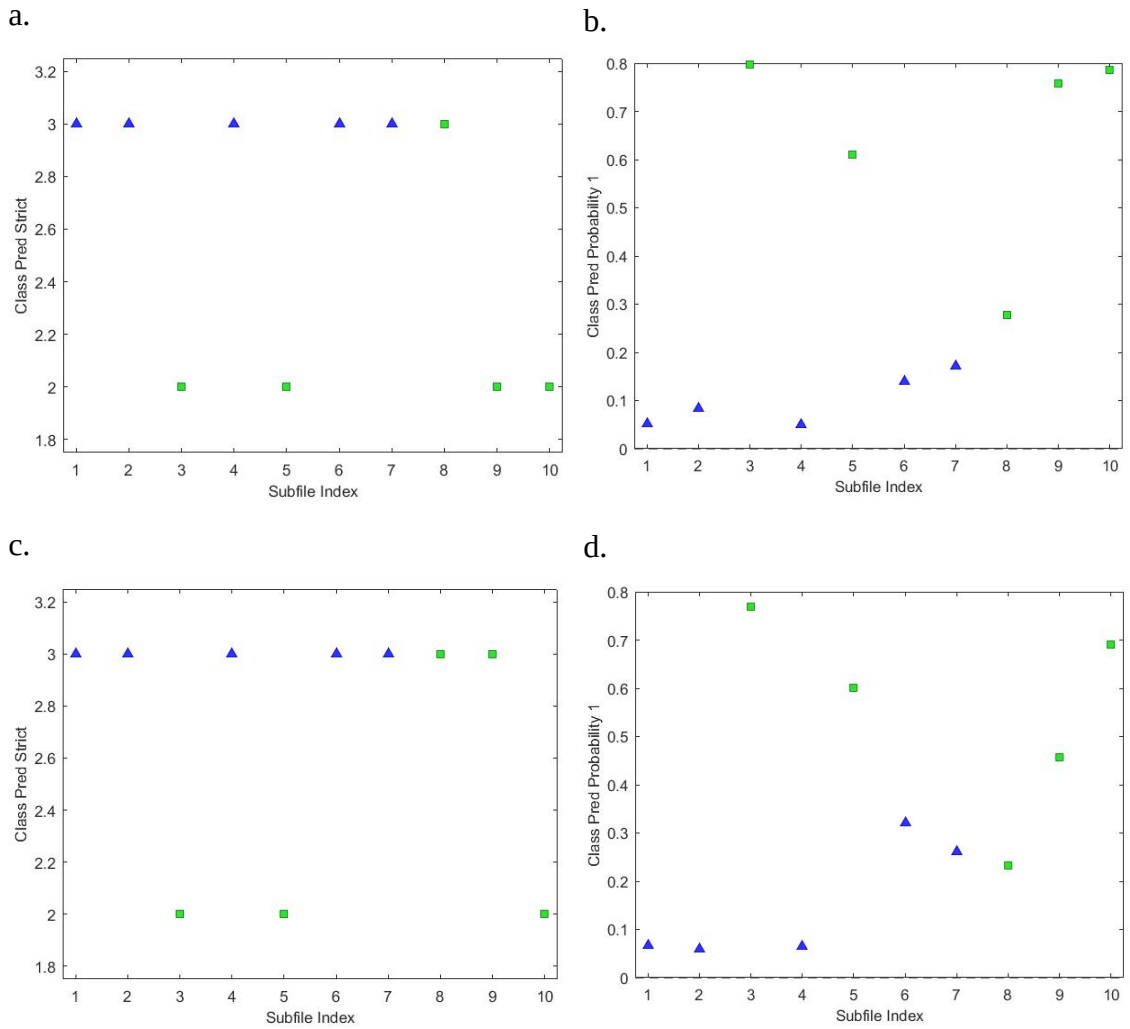


Figure 18 – Scores plots of Models A and A1: **Fig. 18-a.** Strict Class Prediction of Model A; **Fig. 18-b.** Class Prediction Probability of Model A; **Fig.18-c.** Strict Class Prediction of Model A1; **Fig. 18-d.** Class Prediction Probability of Model A1

Figure 19 represents the VIP corresponding to Model A1, where it is possible to observe a peak appearing at 1650 cm^{-1} (proteins), some others between 1540-1500 cm^{-1} (proteins), another one around 1350 cm^{-1} (proteins), a large one between 1160-1145 cm^{-1} (carbohydrates) and another between 965-955 cm-1 (carbohydrates) (see **Annex 13**).

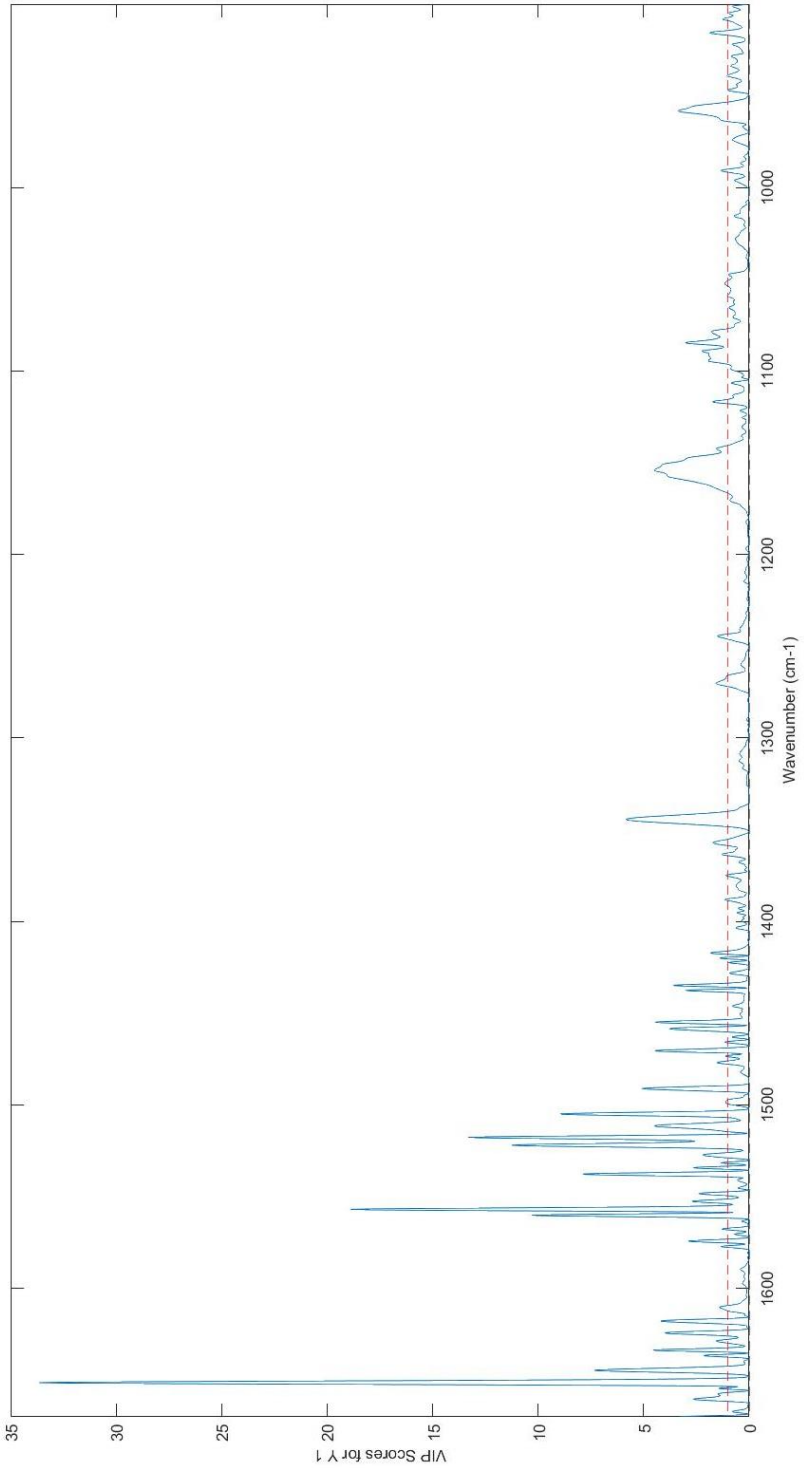


Figure 19 - Importance of wavenumber regions for the A1 model.

About the other created models (B, B1, C and C1), the results are approximately the same to those from A and A1 models, when analyzing Loading VIP scores, with a predominance of proteins and carbohydrates as considered biomolecular alternations (**see Annex 17 and Annex 21**).

In **Annex 14** it is possible to observe that model B1 has a better accuracy than model B, since we have three wrong classified samples (70% of accuracy) against 100% of accuracy in model B1. This can be explained since, observing **Annex 15**, model B is being affected by peculiar peaks before 3500 cm^{-1}, which are not in accordance with the rest of the models. Since, in model B1 (**Annex 16**), spectrum is restricted, these peaks do not influence its results.

About models C and C1, their accuracies are, respectively, 90% and 80%, with 100% of Class Prediction Probability in model C, and less certainty in outcomes in model C1 (see **Annex 18**). In this case, Load VIP scores are very similar to those from the models A and A1, as indicated previously.

With the objective of evaluate the models, we can also explore the cross-validation prevision (confusion metrics). This operates as a simulation of, if each model was applied to the original 46 samples, which would be correctly categorized. Analyzing models confusion tables (see **Annex 22**), it resulted in close results, with an acceptable media value of 67,3% (superior to 2/3) of correct simulations. In **Table 4**, there is a resume about models' accuracies and CV simulations.

Table 4 – PLS-DA models' accuracies and CV simulations resume.

Model	**A**	**A1**	**B**	**B1**	**C**	**C1**
Accuracy (%)	90	80	70	100	90	80
Correct CV simulations (%)	71,7	65,2	67,4	69,6	60,1	69,6

4 Conclusion

With the increasing of cases of CoVid-19, even though vaccination had a main role in slowing down the pandemic, the increasing of affluence to the hospitals led to difficult processes of triage. The hypothesis of using ATR-FTIR coupled with multivariate analysis, as a diagnostic tool, could permit a faster disease severity categorization, throw the presence of higher inflammatory markers.

In this study, we concluded that, despite PCA is not a discriminant method, beyond a better understanding and visualization of spectra behavior, it permitted a division of samples between CoVid-19 positive and negative, with quite close values of accuracy, during different experiments (75,7%, 75,9% and 75,9%). These results could indicate that PCA method is capable of building models that can establish a differentiation of samples, according to their biochemical composition.

In addition, a PLS-DA was developed, using the same spectra as in PCA, which combinations originated six models with quite similar outcomes. The method was able to differentiate the CoVid-19 negative samples from CoVid-19 positive samples, with an accuracy media of 85%. Analyzing the Loadings VIP scores Y, it was possible to observe a predominance of proteins and carbohydrates alternations, being considered the main characteristics to classify the samples. A cross-validation simulation analysis indicated a media of 67.3% of correct simulations. In conclusion, the obtained results suggested that it is possible to distinguish between CoVid-19 positive and negative samples, but further studies are needed to understand how inflammatory process evolution can be quantified and classified, and if the results could be associated/influenced with/by other chronic diseases (for example diabetes or dyslipidemia).

References

1. Tiernan H, Byrne B, Kazarian SG. ATR-FTIR spectroscopy and spectroscopic imaging for the analysis of biopharmaceuticals. Vol. 241, Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy. Elsevier B.V.; 2020.

2. Mohsin GF, Schmitt FJ, Kanzler C, Hoehl A, Hornemann A. PCA-based identification and differentiation of FTIR data from model melanoidins with specific molecular compositions. Food Chemistry. 2019 May 30;281:106–13.

3. Kaznowska E, Depciuch J, Szmuc K, Cebulski J. Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon. Journal of Pharmaceutical and Biomedical Analysis. 2017 Feb 5;134:259–68.

4. Kovács RL, Csontos M, Gyöngyösi S, Elek J, Parditka B, Deák G, et al. Surface characterization of plasma-modified low density polyethylene by attenuated total reflectance fourier-transform infrared (ATR-FTIR) spectroscopy combined with chemometrics. Polymer Testing. 2021 Apr 1;96.

5. Chatchawal P, Wongwattanakul M, Tippayawat P, Kochan K, Jearanaikoon N, Wood BR, et al. Detection of human cholangiocarcinoma markers in serum using infrared spectroscopy. Cancers (Basel). 2021 Oct 2;13(20).

6. Konrad M. Dorling, Matthew J. Baker. Rapid FTIR chemical imaging: highlighting FPA detectors. Trends in Biotechnology [Internet]. 2013 Jun 17; 31(8):437–8. Available from: https://doi.org/10.1016/j.tibtech.2013.05.008.

7. Lovergne L, Bouzy P, Untereiner V, Garnotel R, Baker MJ, Thiéfin G, et al. Biofluid infrared spectro-diagnostics: Pre-analytical considerations for clinical applications. Faraday Discussions. 2016; 187:521–37.

8. Poonprasartporn A, Chan KLA. Live-cell ATR-FTIR spectroscopy as a novel bioanalytical tool for cell glucose metabolism research. Biochimica et Biophysica Acta - Molecular Cell Research. 2021 Jun 1;1868(7).

9. Flora DC, Valle AD, Pereira HABS, Garbieri TF, Buzalaf NR, Reis FN, et al. Quantitative plasma proteomics of survivor and non-survivor COVID-19 patients admitted to hospital unravels potential prognostic biomarkers and therapeutic targets. Available from: https://doi.org/10.1101/2020.12.26.20248855

10. Chen CH, Lin SW, Shen CF, Hsieh KS, Cheng CM. Biomarkers during COVID-19: Mechanisms of Change and Implications for Patient Outcomes. Diagnostics. 2022 Feb 16;12(2):509.

11. 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 - Journal of the American Medical Association. 2020 Mar 17;323(11):1061–9.

12. Guan W jie, Ni Z yi, Hu Y, 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;382(18):1708–20.

13. Du Y, Tu L, Zhu P, Mu M, Wang R, Yang P, et al. Clinical features of 85 fatal cases of COVID-19 from Wuhan: A retrospective observational study. American Journal of Respiratory and Critical Care Medicine. 2020 Jun 1;201(11):1372–9.

14. Samprathi M, Jayashree M. Biomarkers in COVID-19: An Up-To-Date Review. Vol. 8, Frontiers in Pediatrics. Frontiers Media S.A.; 2021.

15. Zhang J jin, Dong X, Cao Y yuan, Yuan Y dong, Yang Y bin, Yan Y qin, et al. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. Allergy: European Journal of Allergy and Clinical Immunology. 2020 Jul 1;75(7):1730–41.

16. Wang F, Nie J, Wang H, Zhao Q, Xiong Y, Deng L, et al. Characteristics of peripheral lymphocyte subset alteration in CoVid-19 pneumonia. Journal of Infectious Diseases. 2020;221(11):1762–9.

17. Yang Y, Shen C, Li J, Yuan J, Wei J, Huang F, et al. Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. Journal of Allergy and Clinical Immunology. 2020 Jul 1;146(1):119-127.e4.

18. Anka AU, Tahir MI, Abubakar SD, Alsabbagh M, Zian Z, Hamedifar H, et al. Coronavirus disease 2019 (COVID-19): An overview of the immunopathology, serological diagnosis and management. Vol. 93, Scandinavian Journal of Immunology. Blackwell Publishing Ltd; 2021.

19. Dorgham K, Quentric P, Gökkaya M, Marot S, Parizot C, Sauce D, et al. Distinct cytokine profiles associated with COVID-19 severity and mortality. Journal of Allergy and Clinical Immunology. 2021 Jun 1;147(6):2098–107.

20. Mravinacova S, Jönsson M, Christ W, Klingström J, Yousef J, Hellström C, et al. A cell-free high throughput assay for assessment of SARS-CoV-2 neutralizing antibodies. New Biotechnology. 2022 Jan 25;66:46–52.

21. Li L, Wu J, Yang L, Wang H, Xu Y, Shen K. Fourier transform infrared spectroscopy: An innovative method for the diagnosis of ovarian cancer. Vol. 13, Cancer Management and Research. Dove Medical Press Ltd; 2021. p. 2389–99.

22. Hands JR, Clemens G, Stables R, Ashton K, Brodbelt A, Davis C, et al. Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. Journal of Neuro-Oncology. 2016 May 1;127(3):463–72.

23. Lin H, Zhang Y, Wang Q, Li B, Huang P, Wang Z. Estimation of the age of human bloodstains under the simulated indoor and outdoor crime scene conditions by ATR-FTIR spectroscopy. Scientific Reports. 2017 Dec 1;7(1).

24. Silva LG, Péres AFS, Freitas DLD, Morais CLM, Martin FL, Crispim JCO, et al. ATR-FTIR spectroscopy in blood plasma combined with multivariate analysis to detect HIV infection in pregnant women. Scientific Reports. 2020 Dec 1;10(1).

25. Santos MCD, Morais CLM, Nascimento YM, Araujo JMG, Lima KMG. Spectroscopy with computational analysis in virological studies: A decade (2006–2016). Vol. 97, TrAC - Trends in Analytical Chemistry. Elsevier B.V.; 2017. p. 244–56.

26. Tufan A, Avanoğlu Güler A, Matucci-Cerinic M. CoVid-19, immune system response, hyperinflammation and repurposinantirheumatic drugs. Vol. 50, Turkish Journal of Medical Sciences. Turkiye Klinikleri; 2020. p. 620–32.

27. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols. 2014;9(8):1771–91.

28. Biancolillo A, Marini F. Chemometric methods for spectroscopy-based pharmaceutical analysis. Vol. 6, Frontiers in Chemistry. Frontiers Media S.A.; 2018.

29. Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemometrics and Intelligent Laboratory Systems. 2012 Aug 1;117:100–14.

30. Miller CE. Chemometrics in Process Analytical Chemistry [Internet]. 2005 [cited 2022 Jun 27]. Available from: https://doi.org/10.1002/9780470988459.ch8

31. Fiolet T, Kherabi Y, MacDonald CJ, Ghosn J, Peiffer-Smadja N. Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review. Vol. 28, Clinical Microbiology and Infection. Elsevier B.V.; 2022. p. 202–21.

32. Dudek T, Knipe DM. Replication-defective viruses as vaccines and vaccine vectors. Vol. 344, Virology. 2006. p. 230–9.

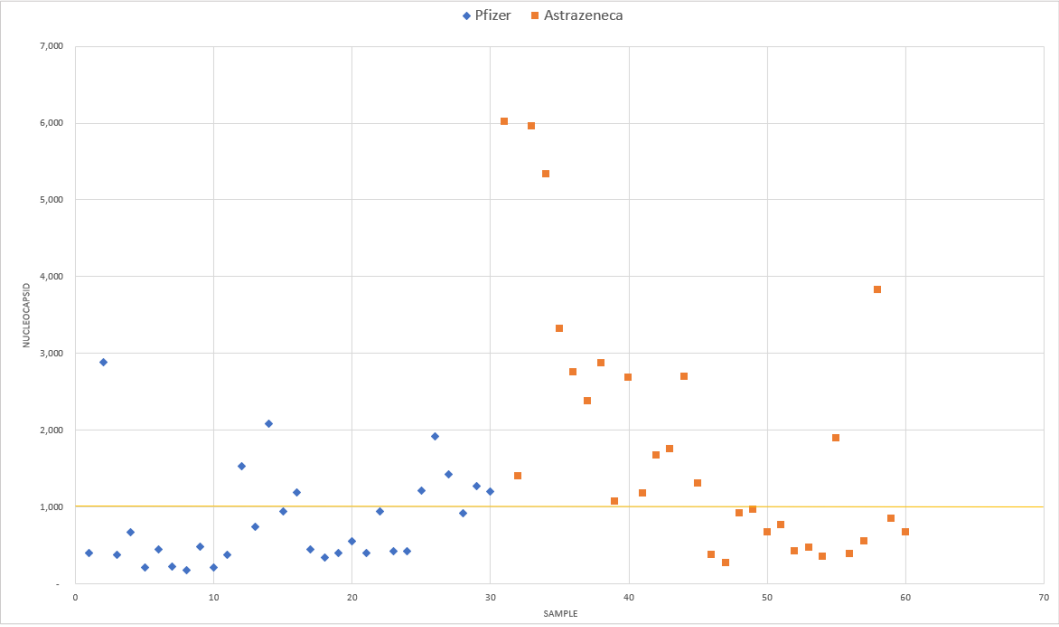
33. Griffin JB, Haddix M, Danza P, Fisher R, Tae ;, Koo H, et al. Morbidity and Mortality Weekly Report SARS-CoV-2 Infections and Hospitalizations Among Persons Aged ≥16 Years, by Vaccination Status-Los Angeles County, California, May 1-July 25, 2021 [Internet]. Available from: https://www.cdc.

Annexes

A1. ATR-FTIR Equipment

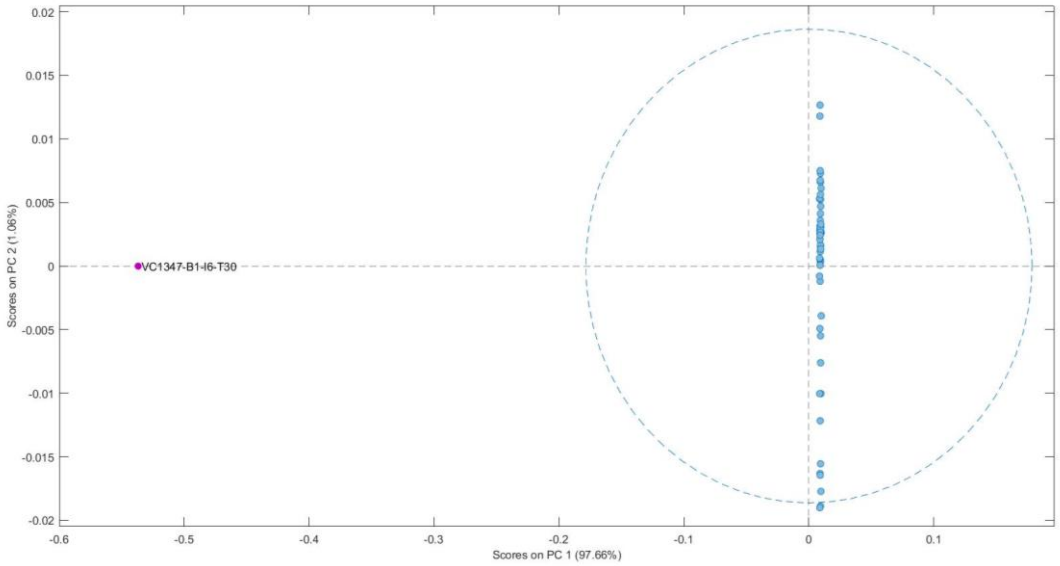
Thermo Scientific™ Nicolet™ IS5 FTIR Spectrometer and diamond crystal *Thermo Scientific™ iD5 ATR* Accessory. Adapted from: https://www.thermofisher.com

A2. Nucleocapsid study samples distribution



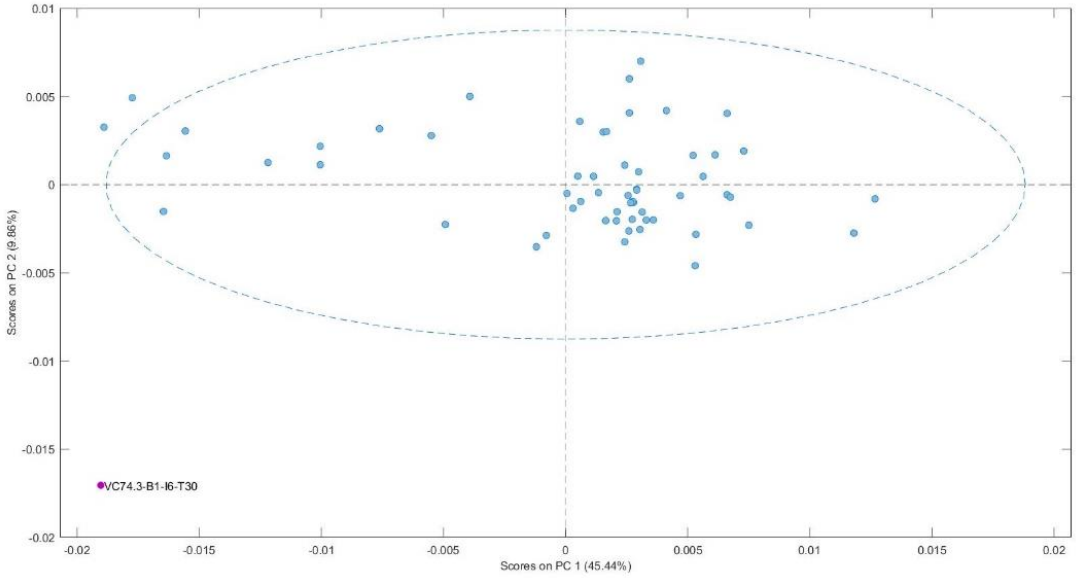
Samples distribution of nucleocapsid study values, separated by vaccines, showing threshold line (orange) corresponding to value 1.

A3. Full spectra PCA analysis showing sample VC1347



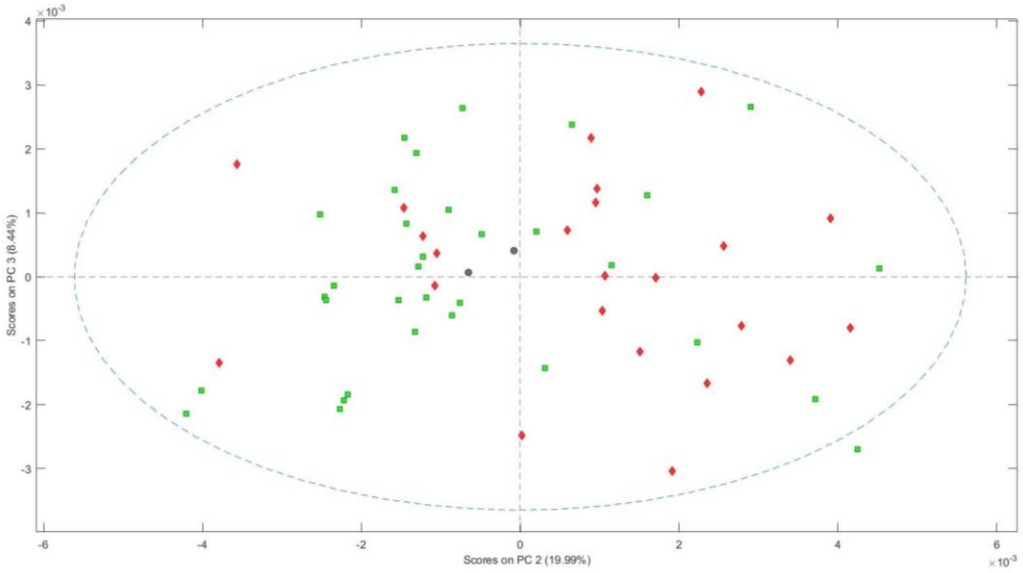
PCA scores plot (PC1 and PC2) showing sample VC1347 influence. There was no wavenumber restriction.

A4. Full spectra PCA analysis showing sample VC74.



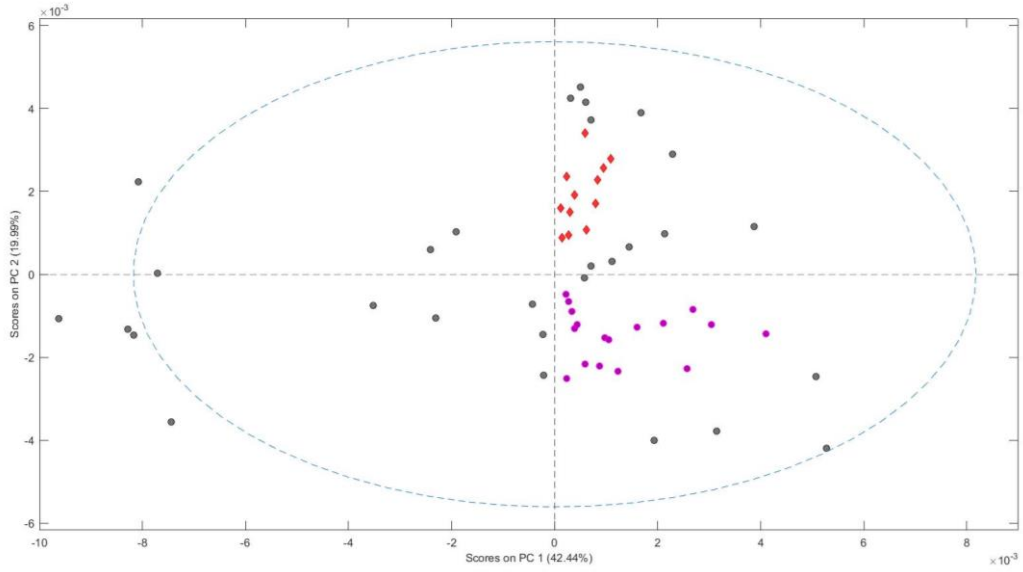
PCA scores plot (PC1 and PC2) showing sample VC74.3 influence. There was no wavenumber restriction.

A5. All spectra PCA scores plot PC2 and PC3



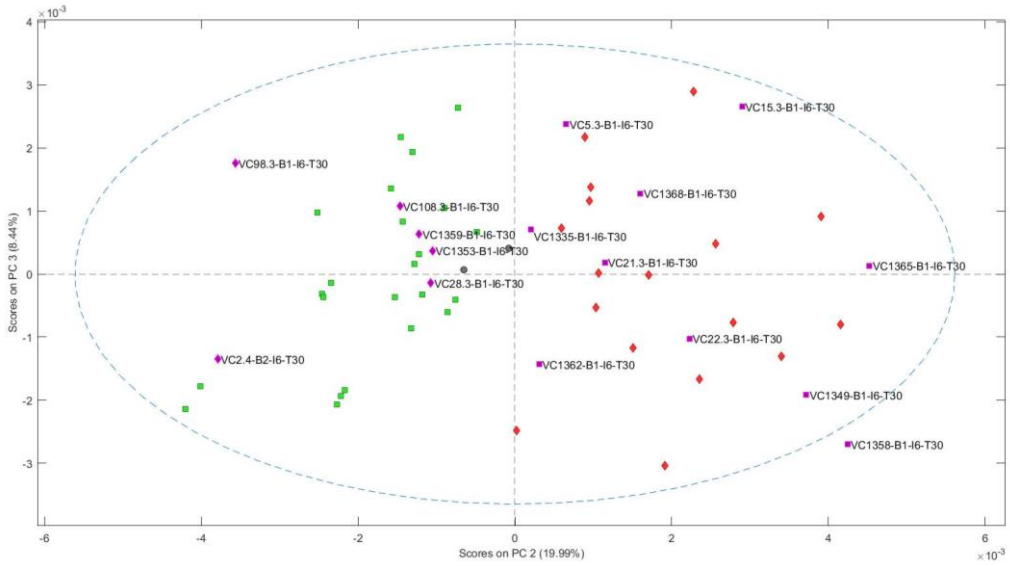
PCA scores plot (PC1 and PC2) showing Covid (red) and non-Covid (green) samples appearing mostly separated according to PC2, with a red area on the right and a green one on the left. Wavenumber was restricted to 1670-900 cm^{-1} .

A6. Hotelling T-Square statistical groups



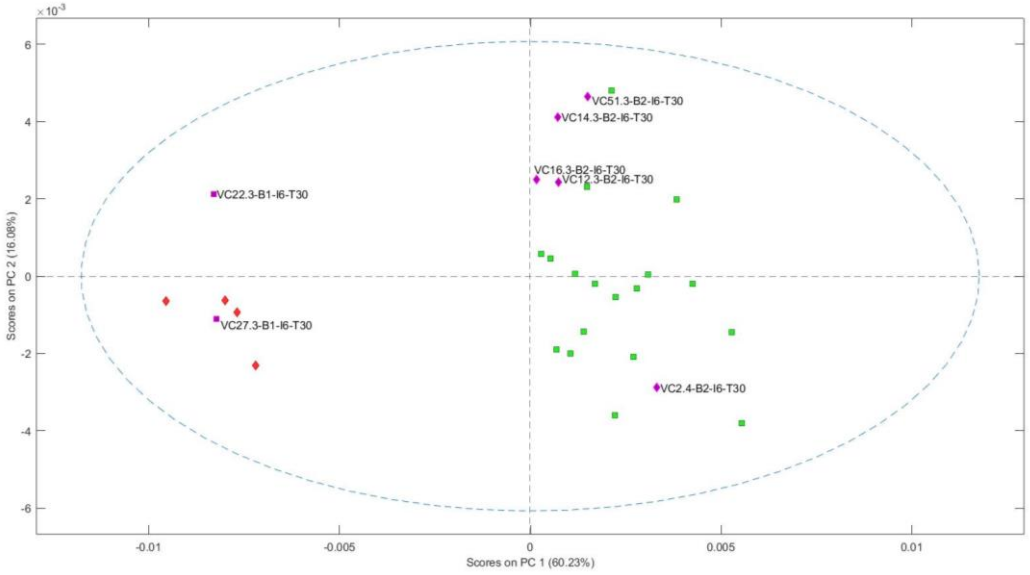
Hotelling T-Square reference group (red) and test group (purple). Wavenumber was restricted to 1670-900 cm-1.

A7. All spectra PCA scores plot PC2 and PC3 (outliers)



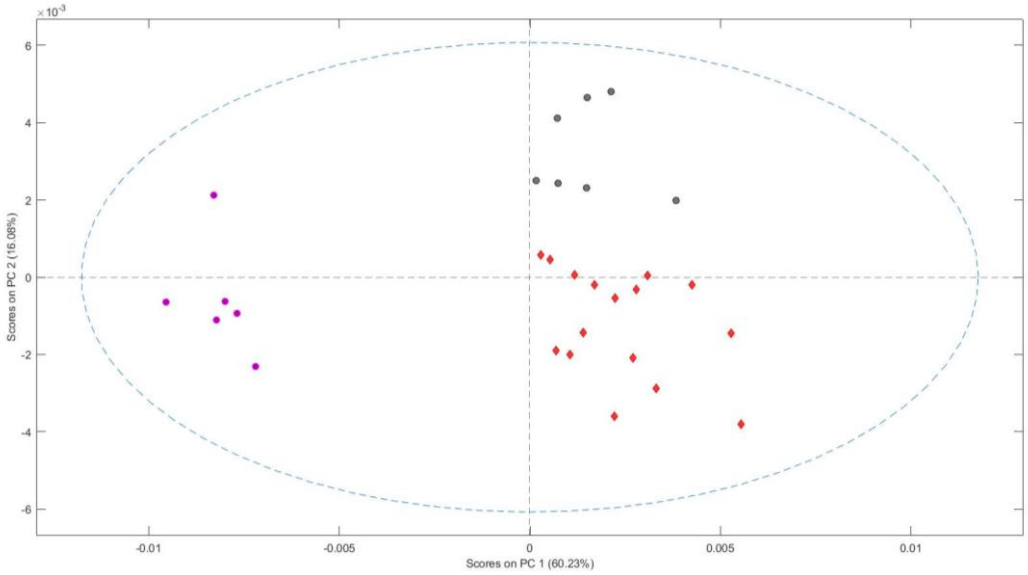
PCA scores plot (PC1 and PC2) showing samples VC98.3, VC2.4, VC108.3, VC1359, VC1353, VC28.3, VC5.3, VC15.3, VC1368, VC21.3, VC22.3, VC1365, VC1349 and VC1358 appearing as outliers, according to CoVid-19 result. Wavenumber was restricted to 1670-900 cm-1.

A8. Pfizer-BioNTech vaccinated PCA scores plot (outliers)



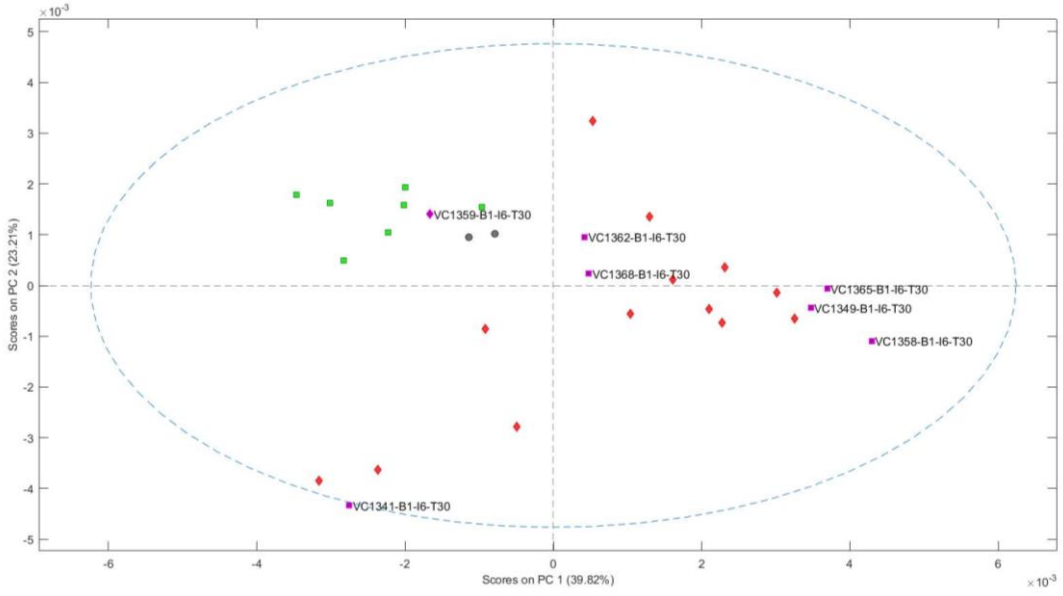
PCA scores plot (PC1 and PC2), from Pfizer-BioNTech vaccinated, showing samples VC2.4, VC12.3, VC14.3, VC16.3, VC22.3 VC27.3 and VC51.3 appearing as outliers, according to CoVid-19 result. Wavenumber was restricted to 1670-900 cm-1.

A9. Hotelling T-Square statistical groups (Pfizer-BioNTech)



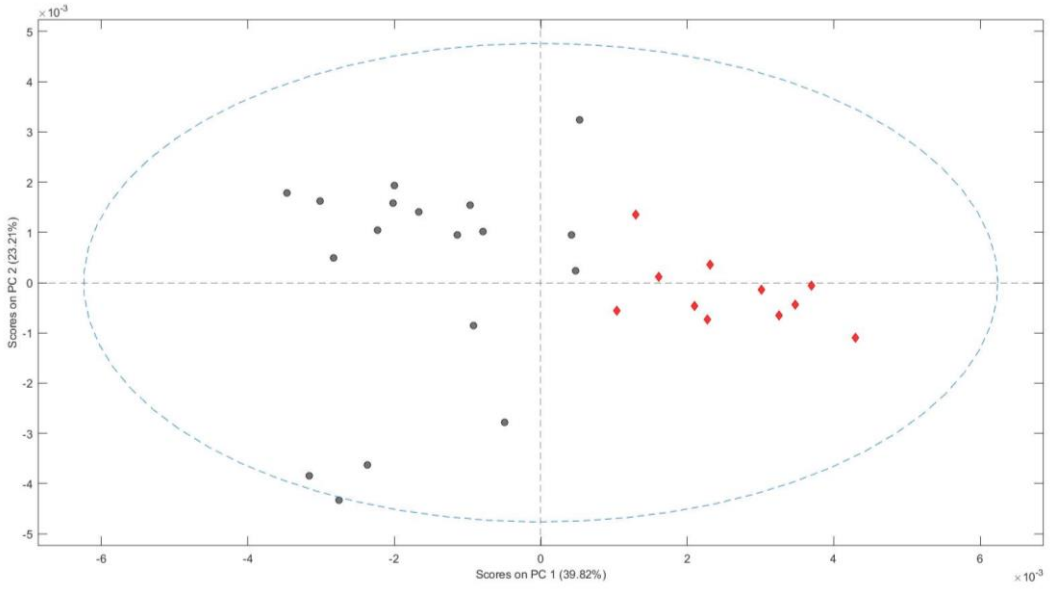
Hotelling T-Square reference group (red) and tests groups (purple and grey). Wavenumber was restricted to 1670-900 cm-1.

A10. AstraZeneca vaccinated PCA scores plot (outliers)



PCA scores plot (PC1 and PC2), from Pfizer-BioNTech vaccinated, showing samples VC1341, VC1349, VC1358, VC1359, VC1362, VC1365, VC1368 appearing as outliers, according to CoVid-19 result. Wavenumber was restricted to 1670-900 cm-1.

A11. Hotelling T-Square statistical groups (Pfizer-BioNTech)



Hotelling T-Square reference group (red) and tests groups (grey). Wavenumber was restricted to 1670-900 cm-1.

A12. PLS-DA models and outcomes resume

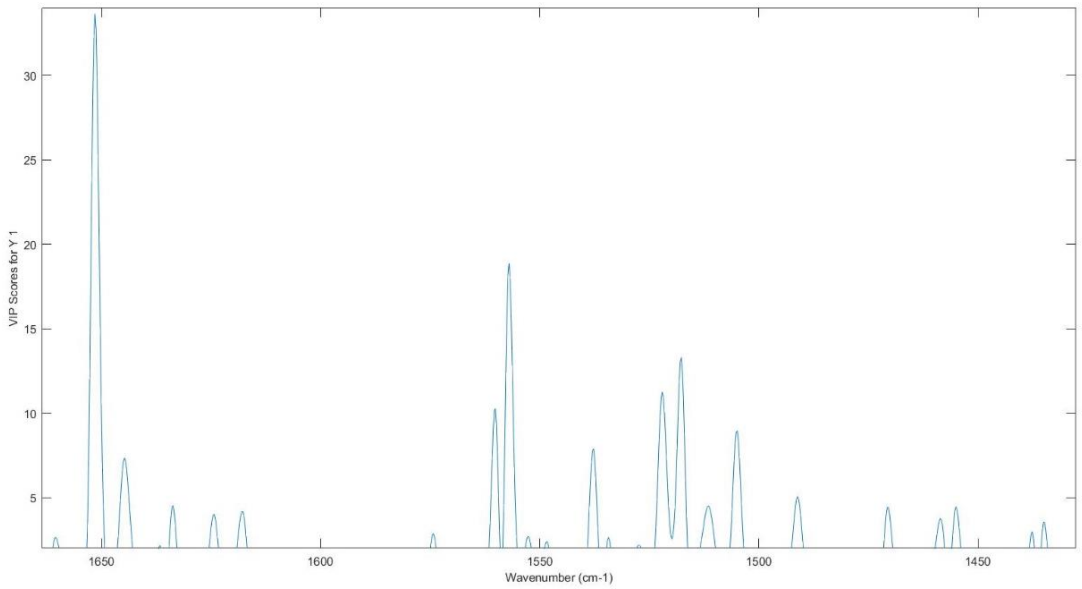
		Test Samples Outcomes					
	Classes	A	A1	B	B1	C	C1
VC1.4_B1_I6_T30	2	2	2				
VC10.3-B1-I6-T30	2	2	2				
VC108.3-B1-I6-T30	1	1	1				
VC11.3-B1-I6-T30	2	2	2				
VC12.3-B2-I6-T30	1	1	1				
VC13.3-B1-I6-T30	2	2	2				
VC1335-B1-I6-T30	2	2	2				
VC1336-B1-I6-T30	1	2	2				
VC1337-B1-I6-T30	1	1	2				
VC1338-B1-I6-T30	1	1	1				
VC1339-B1-I6-T30	1						
VC1340-B1-I6-T30	1						
VC1341-B1-I6-T30	2					1	1
VC1342-B1-I6-T30	1						
VC1344-B1-I6-T30	1					1	1
VC1345-B1-I6-T30	2						
VC1346-B1-I6-T30	2					2	2
VC1349-B1-I6-T30	2						
VC1351-B1-I6-T30	2					2	2
VC1352-B1-I6-T30	2			2	2		
VC1353-B1-I6-T30	1			1	1	1	1
VC1354-B1-I6-T30	1			2	1		
VC1355-B1-I6-T30	1			2	1		
VC1356-B1-I6-T30	1			2	1		
VC1357-B1-I6-T30	1						
VC1358-B1-I6-T30	2						
VC1359-B1-I6-T30	1						
VC1360-B1-I6-T30	2						
VC1362-B1-I6-T30	2						
VC1363-B1-I6-T30	1						
VC1365-B1-I6-T30	2						
VC1367-B1-I6-T30	2						
VC1368-B1-I6-T30	2						
VC14.3-B2-I6-T30	1						
VC15.3-B1-I6-T30	2						
VC16.3-B2-I6-T30	1						

VC17.3-B1-I6-T30	2						
VC18.3-B1-I6-T30	2						
VC19.3-B1-I6-T30	2						
VC2.4-B2-I6-T30	1						
VC20.3-B1-I6-T30	2						
VC21.3-B1-I6-T30	2						
VC22.3-B1-I6-T30	2						
VC26.3-B1-I6-T30	2						
VC27.3-B1-I6-T30	2						
VC28.3-B1-I6-T30	1					1	1
VC3.3_B1_I6_T30	2					2	2
VC4.3-B1-I6-T30	2					2	1
VC45.3-B1-I6-T30	1					1	1
VC5.3-B1-I6-T30	2						
VC51.3-B2-I6-T30	1					1	1
VC6.3-B2-I6-T30	2			2	2		
VC7.3-B2-I6-T30	2			2	2		
VC8.3-B2-I6-T30	2			2	2		
VC9.3-B2-I6-T30	2			2	2		
VC98.3-B1-I6-T30	1			1	1		

PLS-DA models are labeled as A, B, C (no wavenumber restriction) and A1, B1, C1 (1670-900 cm^{-1} wavenumber restriction). Test samples are identified by their corresponding outcome, colored with green (outcome agreeing with its respective class) or red (distinct outcome and classification).

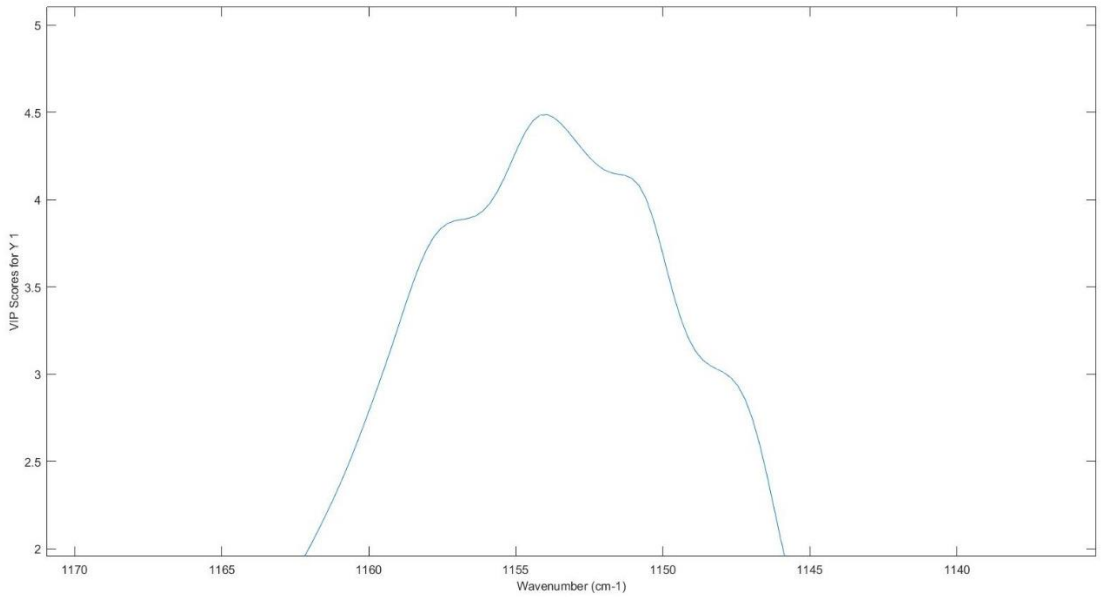
A13. PLS-DA Loading VIP scores Zoom from model A1

a.



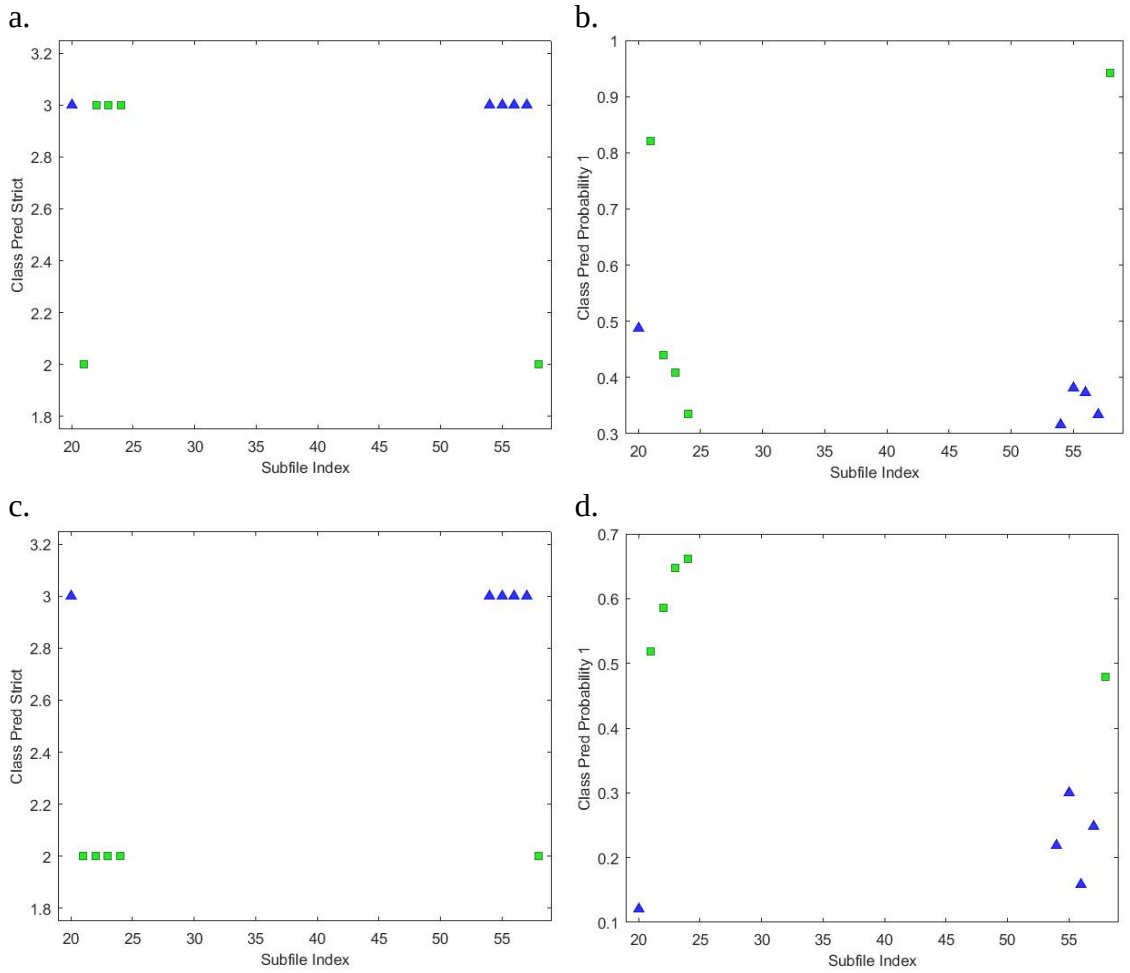
PLS-DA Loading VIP score Zoom of wavenumbers bettween 1660-1450 cm^{-1} (proteins) from model A1.

b.



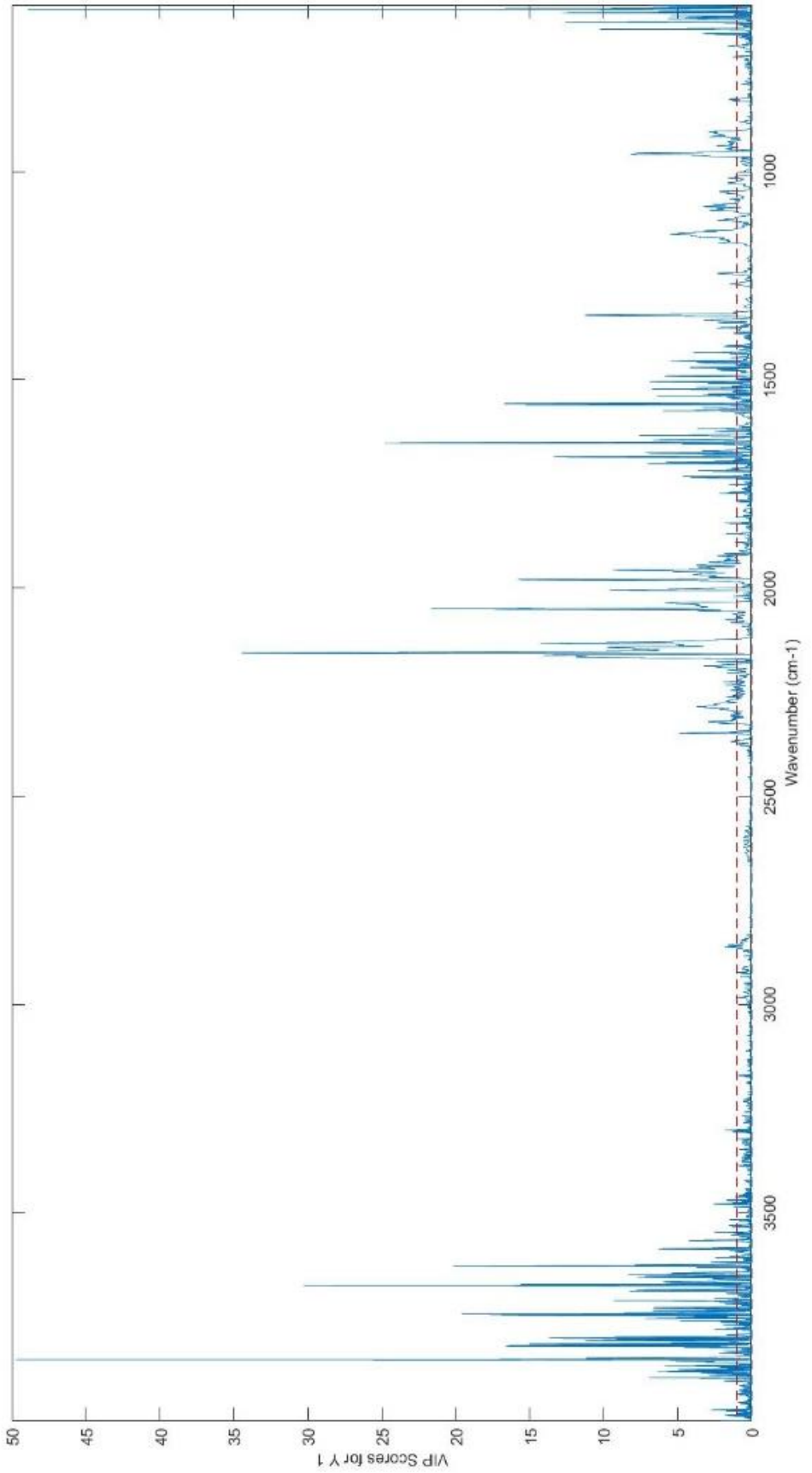
PLS-DA Loading VIP score Zoom of wavenumbers bettween 1165-1145 cm^{-1} (carbohydrates) from model A1.

A14. PLS-DA scores from model B and B1

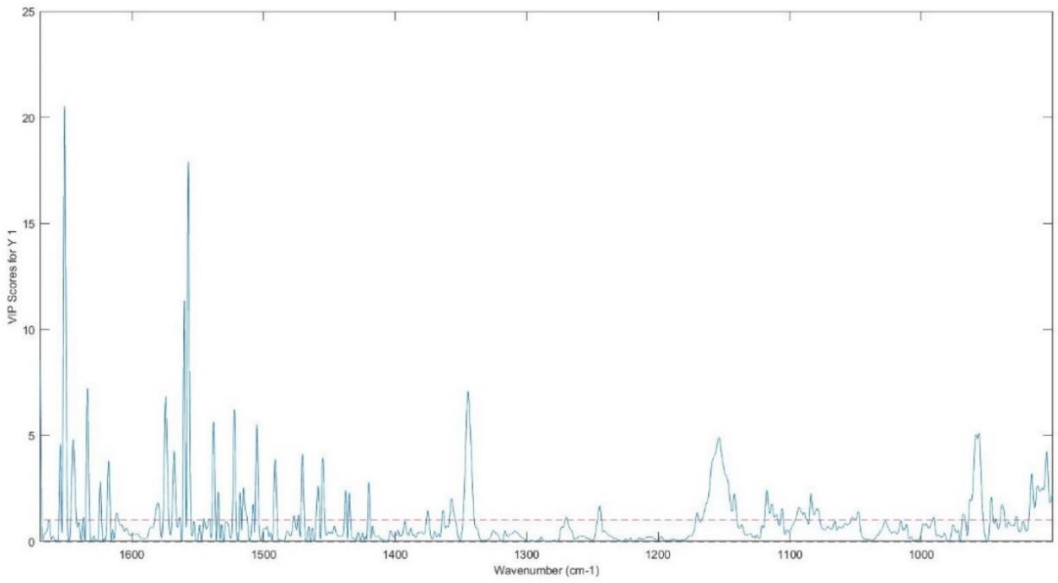


Scores plots of Models B and B1: **A14-a.** Strict Class Prediction of Model B; **A14-b.** Class Prediction Probability of Model B; **A14-c.** Strict Class Prediction of Model B1; **A14-d.** Class Prediction Probability of Model B1

A15. PLS-DA Loading VIP scores from model B

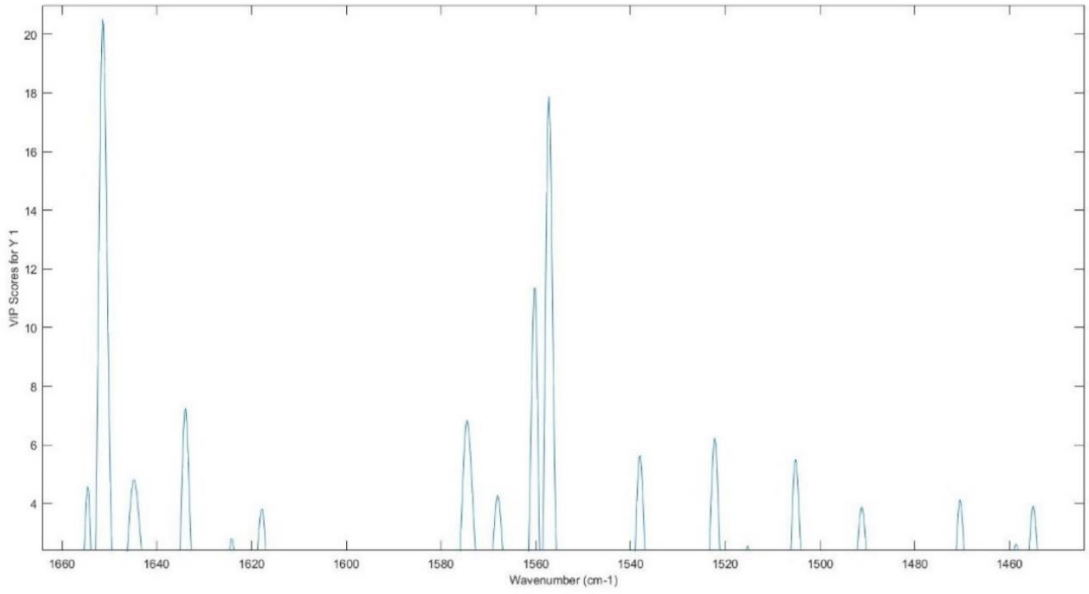


A16. PLS-DA Loading VIP scores from model B1



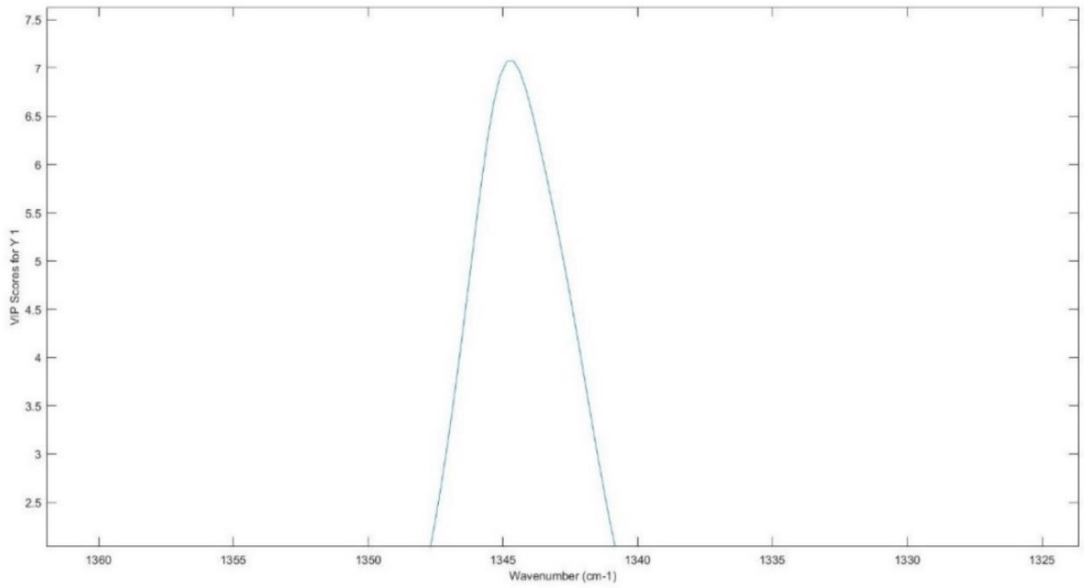
A17. PLS-DA Loading VIP scores Zoom from model B1

a.



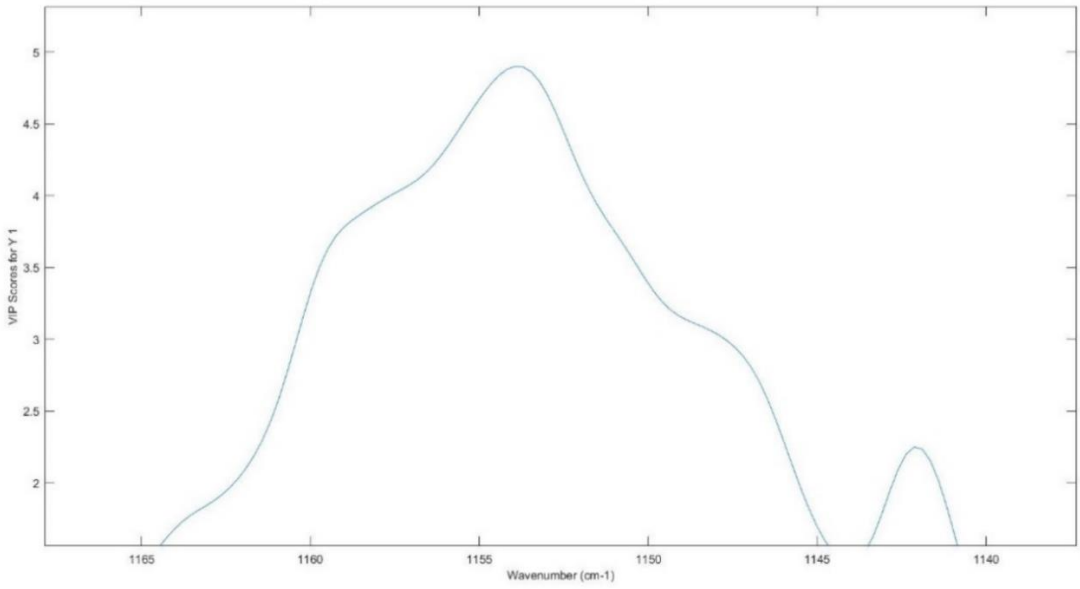
PLS-DA Loading VIP score Zoom of wavenumbers bettween 1660-1450 cm^{-1} (proteins) from model B1.

b.



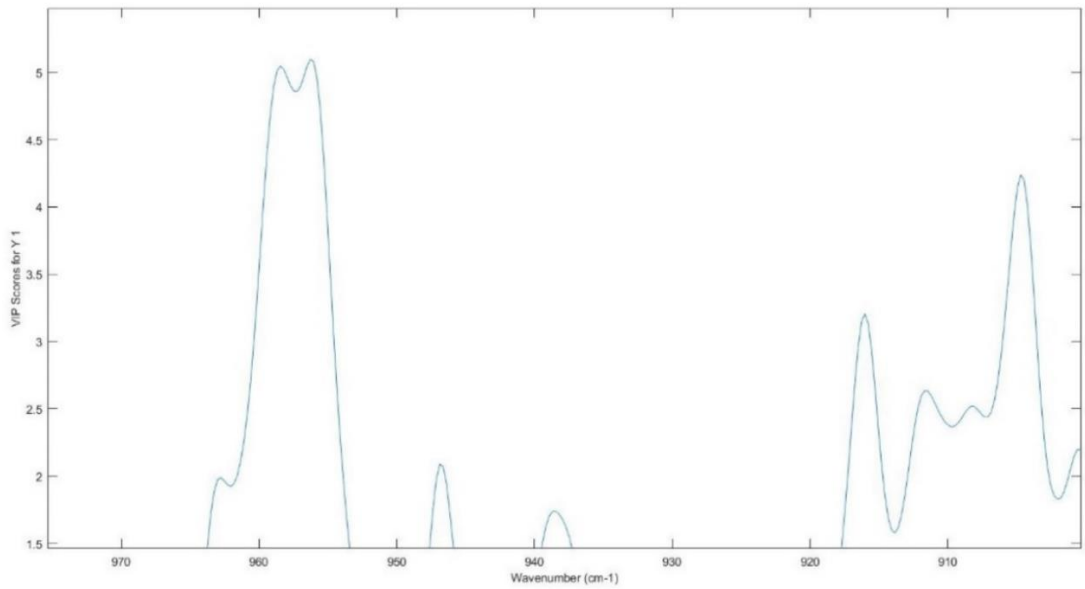
PLS-DA Loading VIP score Zoom of wavenumbers bettween 1350-1340 cm^{-1} (proteins) from model B1.

c.



PLS-DA Loading VIP score Zoom of wavenumbers bettween 1165-1140 cm^{-1} (carbohydrates) from model B1

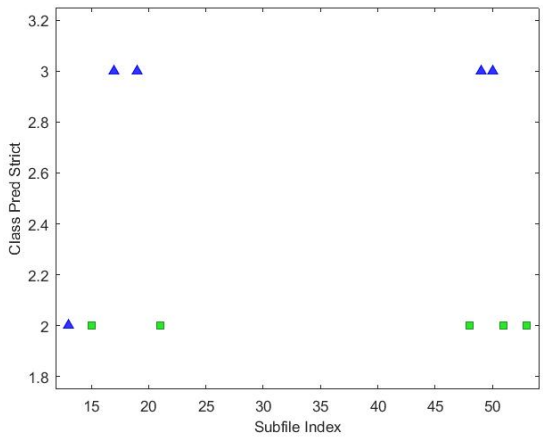
d.



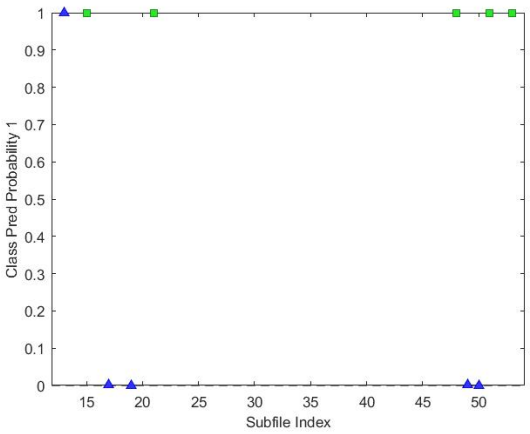
PLS-DA Loading VIP score Zoom of wavenumbers bettween 965-900 cm^{-1} (carbohydrates) from model B1.

A18. PLS-DA scores from model C and C1

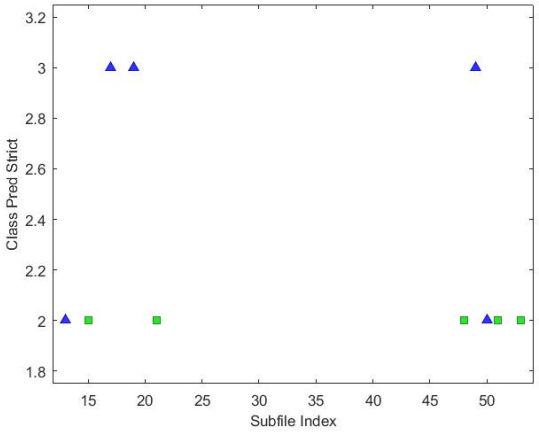
a.



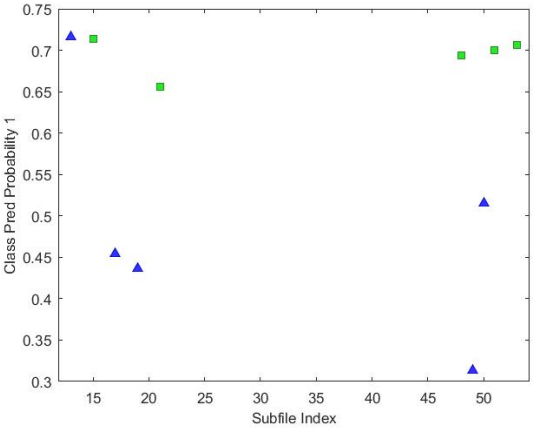
b.



c.

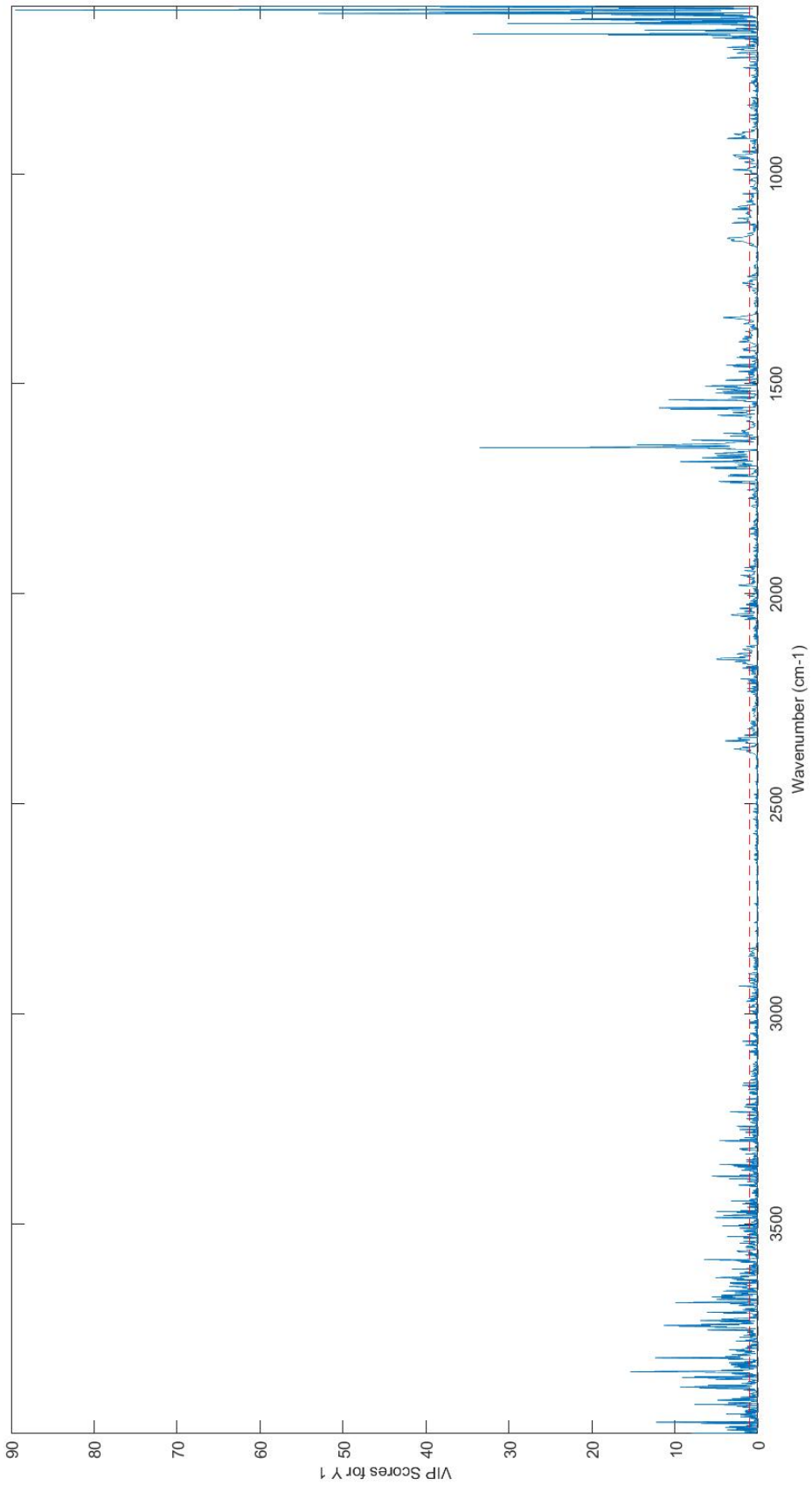


d.

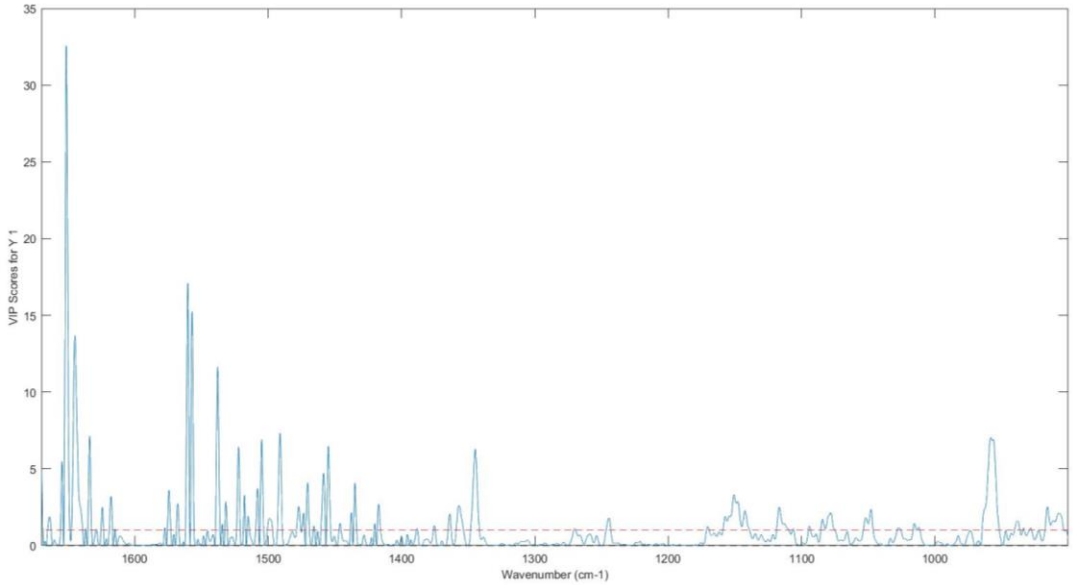


Scores plots of Models C and C1: **A18-a.** Strict Class Prediction of Model C; **A18-b.** Class Prediction Probability of Model C; **A18-c.** Strict Class Prediction of Model C1; **A18-d.** Class Prediction Probability of Model C1

A19. PLS-DA Loading VIP scores from model C

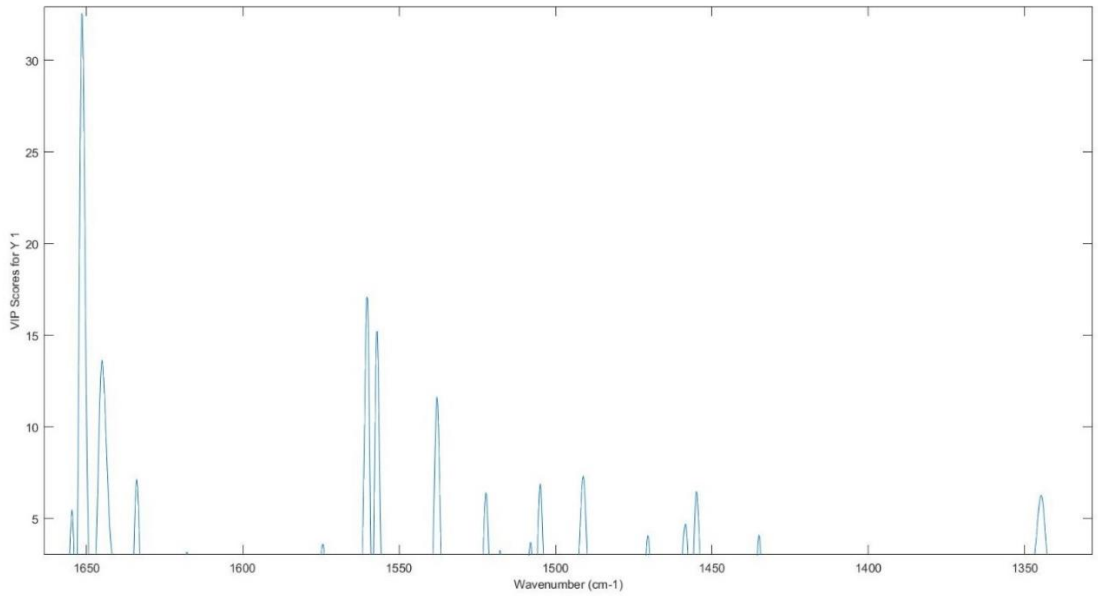


A20. PLS-DA Loading VIP score from model C1



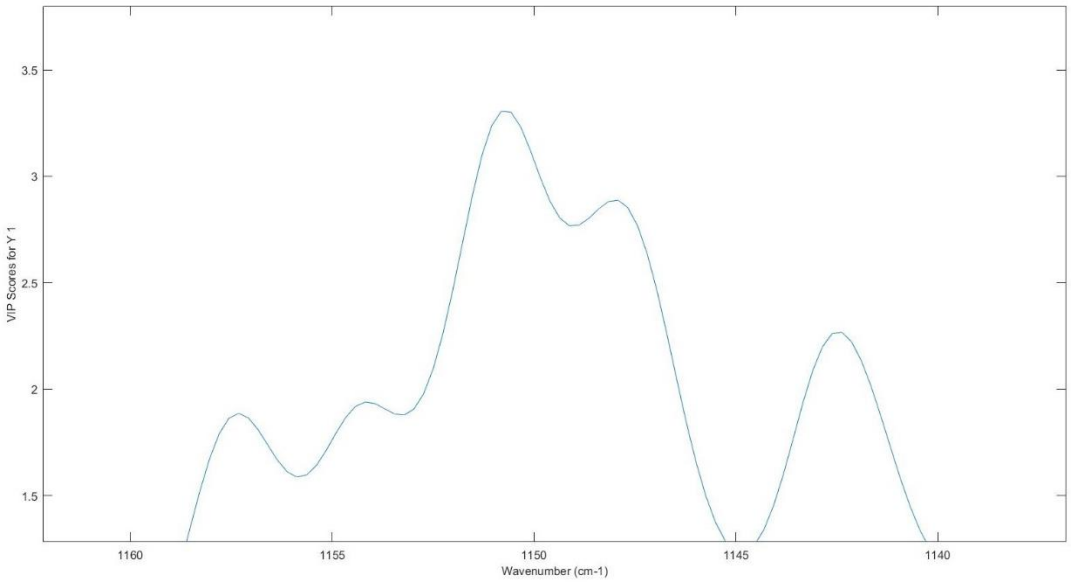
A21. PLS-DA Loading VIP scores Zoom from model C1

a.



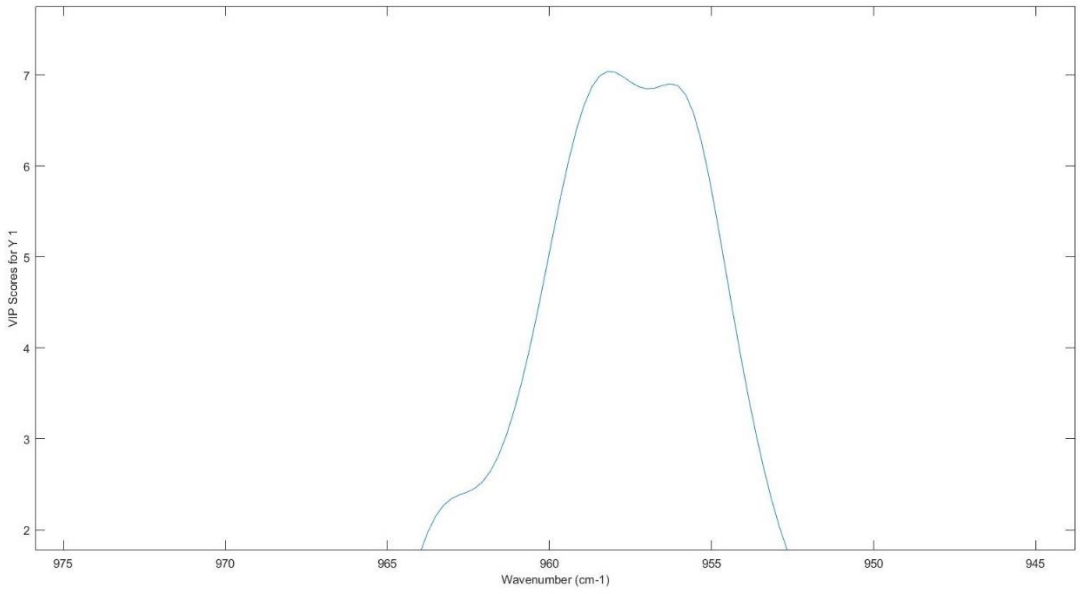
PLS-DA Loading VIP score Zoom of wavenumbers bettween 1660-1340 cm^{-1} (proteins) from model C1.

b.



PLS-DA Loading VIP score Zoom of wavenumbers bettween 1160-1140 cm^{-1} (carbohydrates) from model C1.

c.



PLS-DA Loading VIP score Zoom of wavenumbers bettween 965-950 cm^{-1} (carbohydrates) from model B1.

A22. PLS-DA Cross-Validation simulations

Confusion Table A (CV):		
	Actual Class	
	1	2
Predicted as 1	11	9
Predicted as 2	7	19

Confusion Table A1 (CV):		
	Actual Class	
	1	2
Predicted as 1	12	10
Predicted as 2	6	18

Confusion Table B (CV):		
	Actual Class	
	1	2
Predicted as 1	6	3
Predicted as 2	12	25

Confusion Table B1 (CV):		
	Actual Class	
	1	2
Predicted as 1	13	9
Predicted as 2	5	19

Confusion Table C (CV):		
	Actual Class	
	1	2
Predicted as 1	11	11
Predicted as 2	7	17

Confusion Table C1 (CV):		
	Actual Class	
	1	2
Predicted as 1	15	11
Predicted as 2	3	17

Simulation of corresponding actual and predicted classes, during A, A1, B, B1, C and C1 model cross-validation.