Instituto Politécnico de Lisboa

Instituto Superior de Engenharia de Lisboa

Escola Superior de Tecnologia de Saúde de Lisboa

Impact of Epigallocatechin-3-gallate (EGCG) on the molecular profile of plasma and serum

Rúben Alexandre Dinis Araújo

Thesis to obtain the Master Degree in

Biomedical Engineering

Definitive version

Supervisors:

Dr. Cecília Ribeiro da Cruz Calado (ISEL)

Dr. Edna Ribeiro (ESTeSL)

December 2019

Instituto Politécnico de Lisboa

Instituto Superior de Engenharia de Lisboa

Escola Superior de Tecnologia de Saúde de Lisboa

Impact of Epigallocatechin-3-gallate (EGCG) on the molecular profile of plasma and serum

Rúben Alexandre Dinis Araújo

Thesis to obtain the Master Degree in

Biomedical Engineering

Supervisors:

Dr. Cecília Ribeiro da Cruz Calado (ISEL)

Dr. Edna Ribeiro (ESTeSL)

Examination Committee

Chairperson:

Dr. Manuel Matos (ISEL)

Members of the Committee:

Dr. Miguel Brito (ESTeSL)

Dr. Luís Filipe Nunes Bento (NOVA Medical School – Universidade Nova de Lisboa)

December 2019

[This page is intentionally left blank]

"Science may have found a cure for most evils;
but it has found no remedy for the worst of them all
– the apathy of human beings."

Hellen Keller

[This page is intentionally left blank]

Impact of Epigallocatechin-3-gallate (EGCG) on the molecular profile of plasma and serum

Rúben Alexandre Dinis Araújo

This work was supported by *Instituto Politécnico de Lisboa* under the grant IDI&CA/IPL/2018/RenalProg/ISEL. The present work was partially conducted in the Engineering & Health Laboratory through a collaboration protocol established between *Universidade Católica Portuguesa* and *Instituto Politécnico de Lisboa*.

[This page is intentionally left blank]

Acknowledgments

To be fair to everyone, I would like to thank the people that contributed to this thesis in any way possible, by order of appearance in my new life, ever since I came back to Portugal, after the many years spent in Norway.

As such, it is a *lapalissade* to refer Professor Cecília Calado as the first in my list. Our path crossed from the moment I decided to come back to Portugal and attempt a drastic direction in my life. It was first through a Skype meeting of two people, thousands of kilometres away from each other, that I was convinced to try a new Master degree in ISEL that had piqued my interest: Biomedical Engineering. From the first seminar class, to being able to work with her and her team, it has been a pleasure and an honour to learn under her umbrella and her colleagues. Her objective critique and straight forward thinking made me feel at home and allowed to perform at a level that, to be perfectly honest, I wouldn't otherwise have bothered with. A teacher, a colleague and a friend. This is what I am taking from this relationship.

I extend my gratitude to my lab colleagues, from the oldest ones which have since left (Joana, Bernardo) and went on with their lives, to the ones that are there now (thank you Filipa Pires for the enthusiasm into themes that at times would be just plain boring, I promise to explain to you our labs' secret word after you graduate), and to the ones that will come next. Thank you. Of these I would like to extend special thanks to those that have helped me find my way around a lab and with my thesis. Thank you Maria João Pereira for helping out with the ins and outs of a laboratory and the initial entry into OPUS and The Unscrambler® X and a special thank you for Professor Edna Ribeiro for supplying with the raw biomaterials that allowed the making of this work and for Helder da Paz, a lab colleague second but first and foremost, a friend and a brother, for processing the samples and allowing this work to exist. We miss your cheerfulness and singing echoing throughout the laboratory. But undeniably, the one who takes the biggest slice of the cake is Luís Ramalhete. Colleague, teacher, friend for life and go to person for the heavy hitting questions about biology and machine learning. It was fun discussing the inner workings of the many subjects I would dare not thread alone.

A special acknowledgement goes to the 1st class of Biomedical Engineering students in ISEL, for giving me hope in future generations and encouraging me to be better for the ones that come after us. It will be a pleasure seeing you all grow, both as students, IEEE student members and above all as human beings.

Finally, a heartfelt note to all my closest friends and loved ones. You are not that many. You are not even that close. Not only land but oceans and time itself separate us now. Regardless, know that you are close to my heart. My wish now is that after this new page in my life finally unfolds, I will once again regain the reigns in my life and once more be able to walk with you again. I miss you.

[This page is intentionally left blank]

Abstract

Background and Goals: Epigallocatechin-3-gallate (EGCG) is the major catechin present in green tea and it is known to display diverse biological activities as antioxidation, antiinflammation, antiproliferation, antimicrobial, antiviral, among others. The present work aimed to evaluate the impact of consumption of EGCG on the molecular profile of human serum and plasma.

Methods: The effect of the consumption of a daily intact of 225 mg for 90 days of EGCG on healthy human volunteers (n = 30), on the plasma and serum molecular profile was evaluated by mid-infrared spectroscopy (MIRS). A method to search for biomarkers in human plasma and serum was developed based on MIRS and machine learning methods.

Results: It was observed through different unsupervised pattern search methods and classification supervised methods, e.g. Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR), Discriminant Analysis (DA), Hierarchical Cluster Analysis (HCA) that, both plasma and serum samples presented a significantly different molecular profile after the 90 days of EGCG consumption. Based on loadings, the regions of the spectra with the most impact into cluster separation between 90 days were analysed and it was observed that EGCG consumption affected the profile of major molecules as proteins and lipids. Diverse absorbance ratios were identified as being statistically different ($p < 0.01$) after EGCG consumption, revealing a high impact of EGCG on human general metabolism.

Conclusions: MIRS enabled to monitor the drastic change of the molecular profile of serum and plasma after 90 days of EGCG consumption. MIRS allowed to attain the molecular profile of the sample in a sensitive and specific mode, but also in an economic, simple, rapid and high throughput way. The technique, when combined with automatic methods of pattern recognition, classification and biomarkers search, results in a highly innovative and promising method to acquire information in large-scale epidemiological studies towards a better understanding of the *in vivo* effect of EGCG.

Keywords: Green tea, epigallocatechin-gallate, biomarkers, machine learning

Resumo

Enquadramento e Objectivos: A epigalocatequina-3-galato (EGCG) é a catequina principal presente no chá verde, apresentando diversas propriedades biológicas como antioxidação, anti-inflamação, anti-proliferação, antimicrobianas, anti-virais, entre outras. O presente trabalho teve como objetivo estudar o impacto do consumo de EGCG no perfil molecular do soro e plasma humano.

Métodos: Foi estudado o efeito do consumo diário de 225 mg durante 90 dias de EGCG, em voluntários humanos saudáveis (n = 30), no perfil molecular do plasma e soro, adquirido por espectroscopia de infravermelho médio (do inglês MIRS – *mid infrared spectroscopy*). Foram estudados biomarcadores, através do uso de MIRS associado a métodos de aprendizagem automática.

Resultados: Foi observado através de diferentes métodos de reconhecimento de padrões não supervisionados e de classificação supervisionados, como por exemplo de Análise de Componentes Principais (do inglês PCA – *Principal Component Analysis*), Regressão Parcial de Mínimos Quadrados (do inglês PLSR, *Partial Least Squares Regression*), Análise Discriminate (do inglês DA – *Discriminant Analysis*), Análise de Agrupamentos Hierárquicos (do inglês HCA – *Hierarchical Cluster Analysis*), que ambas as amostras de plasma e soro apresentam um perfil molecular significativamente diferente após 90 dias de consumo de EGCG. Com base nos *loadings*, foram analisadas as regiões do espectro que mais contribuíram para a separação dos grupos antes e após os 90 dias de consumo, tendo sido observado que o consumo de EGCG afetou o perfil de moléculas como proteínas e lípidos. Foram identificadas diversas razões de absorvância como estatisticamente diferentes ($p < 0.01$) após consumo de EGCG, revelando um elevado impacto do consumo de EGCG no metabolismo geral do ser humano.

Conclusões: A MIRS permitiu monitorizar as elevadas alterações moleculares observadas no plasma e soro após 90 dias de consumo de EGCG. A MIRS permitiu obter o perfil molecular das amostras de uma forma muito sensível e específica, mas também económica, simples, rápida e de alto débito. Esta técnica, quando combinada com métodos de reconhecimento de padrões, de classificação e de pesquisa de biomarcadores, resulta num método inovador e bastante promissor na aquisição de informação em estudos epidemiológicos de larga escala e contribui para uma melhor compreensão do efeito *in vivo* do EGCG.

Palavras-chave: Chá verde, epigalocatequina-galato, biomarcadores, aprendizagem automática

Relevant publications in the field

Proceeding

Araujo RAD, Ramalhete LM, Ribeiro E, Calado CRC. Effect of consumption of green tea extracts on the plasma molecular signature. In: 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG). IEEE; 2019:1-4. doi:10.1109/ENBENG.2019.8692524

[This page is intentionally left blank]

Table of contents

List of Figures

List of Tables

List of Abbreviations

AIMA	Automated iterative moving average
airPLS	Adaptive iteratively reweighted penalized least squares
ANN	Artificial neural network
AUC	Area under the curve
BC	Baseline correction
BEST	Biomarkers, endpoints and other tools
CA	Classification accuracy
CNS	Central nervous system
DA	Discriminant analysis
DT	De-trending
EC	Epicatechin
ECG	Epicateching-3-gallate
EDA	Exploratory data analysis
EDTA	Ethylenediaminetetraacetic acid
EGC	Epigallocatechin
EGCG	Epigallocatechin-3-gallate
EMSC	Extended multiplicative scatter correction
FDA	Food and drug administration
FIR	Far infrared
FTIRS	Fourier transform infrared spectroscopy
GABA	Gamma aminobutyric acid
GC	Gas chromatography
GTE	Green tea extract
HbF	Fetal haemoglobin
HC	Hierarchical cluster
HCA	Hierarchical cluster analysis
HPLC	High Performance Liquid Chromatography
IA	Iterative average
IR	Infrared
LDA	Linear discriminant analysis
LOO	Leave one out
MIR	Mid infrared
MIRS	Mid infrared spectroscopy
ML	Machine learning
MLR	Multiple linear regression
MPLS	Morphological weighted penalized least squares
MSC	Multiplicative scatter correction
MVA	Multivariate analysis
NIH	National institutes of health
NIR	Near infrared
NLCE	Nanostructured lipid carriers
PC	Principal components
PCA	Principal component analysis
PCR	Principal components regression
PLSR	Partial Least Squares Regression
RMSECV	Root mean square error of cross validation
ROC	Reactive oxygen species
ROC	Receiver operation characteristic
SEC	Standard error of calibration
SG	Savitzky-Golay
SNV	Standard normal variate
SVM	Support vector machine
TOXNET	Toxicology data network of the USA National Institute of Health

[This page is intentionally left blank]

Chapter 1: Thesis Objectives and Work Structure

There is an increasing need to develop rapid and precise tools that allow, among other things, the ability to detect early signs of human pathologies, enhance clinical diagnostics, accelerate clinical trials and offer reliable and meaningful methods to analyse highly complex biological systems. General omics techniques and Fourier-transform infrared spectroscopy (FTIRS) are two such tools [1]. The main goal of the present thesis was to develop a rapid, reproducible, sensitive and specific methodology for the discovery of biomarkers in plasma and serum enabling to evaluate the physiological effect of 90 days of a daily ingestion of 225 mg of EGCG extract. It was also aimed to develop the methodology based on plasma and serum FTIR spectra associated to a comprehensive machine learning workflow. To achieve that, the following were also aimed:

1. Identify and quantify the differences in the clinical variables for plasma and serum before and after EGCG consumption;

2. Develop a database based on FTIR spectra of serum and plasma of 30 healthy individuals taken before the beginning of daily intake of EGCG extract and at the end of 90 days of consumption;

3. Develop a workflow of supervised and unsupervised learning methods in order to automatically conduct the FTIR database pre-processing, cluster separation, class prediction and biomarker discovery;

4. Identify the spectral bands in which EGCG had a more significant impact on the general metabolism and suggest possible biomarkers for both plasma and serum;

5. Develop in parallel to the work conducted in the two previous points, automated machine learning techniques to biomarkers discovery as highlighted in as highlighted in Figure 1.1.1.

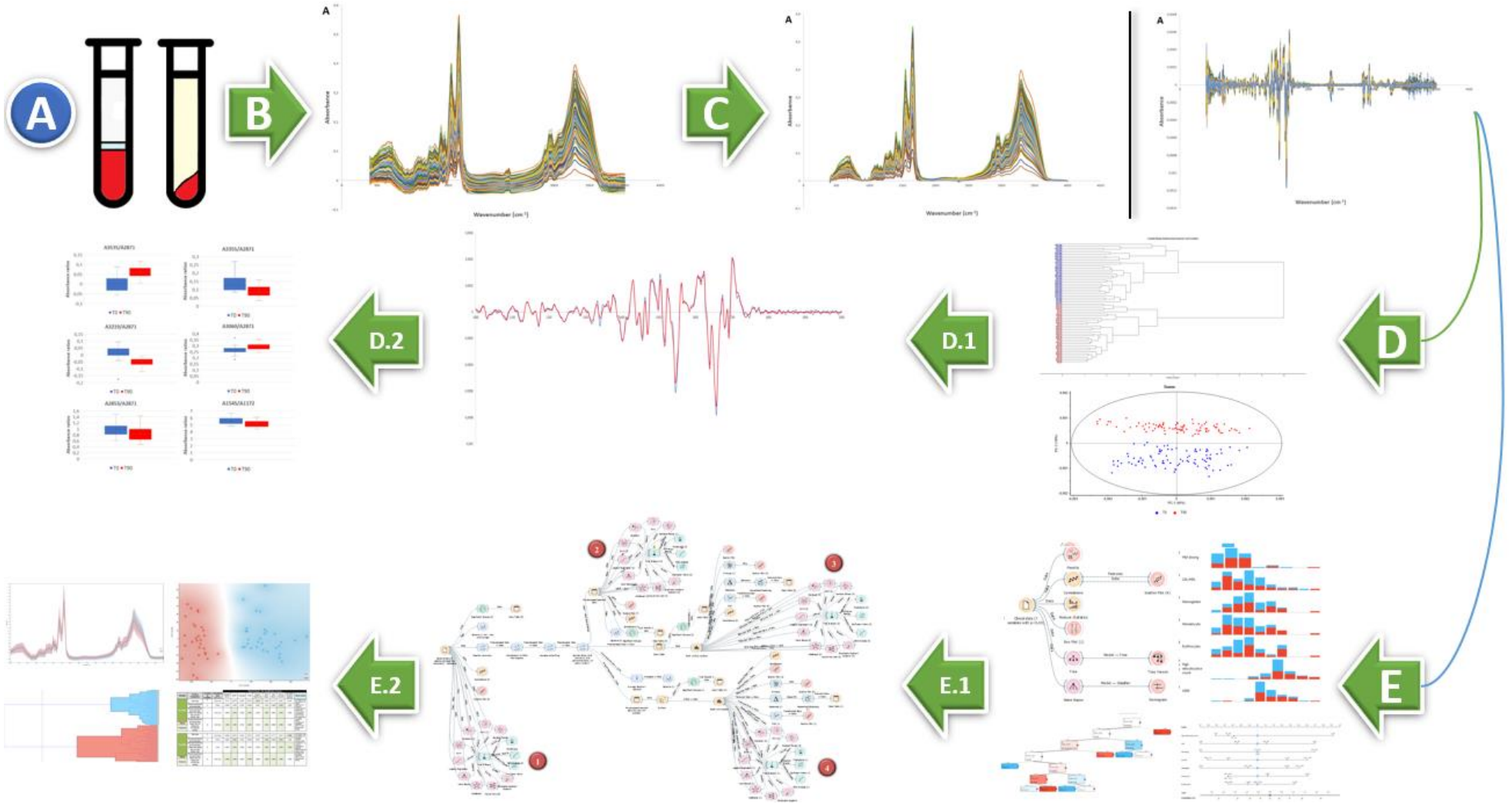


*Figure 1.1.1. Simplified thought process behind the work's workflow and biomarker discovery. (**A**) Biofluids of plasma and serum are acquired in a clinical laboratory environment. (**B**)Biological samples are then processed by vibrational spectroscopy techniques (FTIRS). (**C**) The acquired spectra are submitted to pre-processing by atmospheric correction, baseline correction, derivatives and others. The workflow is then transformed into a parallel process. (**D**) The use of multivariate analysis techniques, as PCA score plots and HCA, are used to identify different clusters in the studied population (n=30). (**D.1.**) Through the use of loadings and spectra differences between the different groups (T0 and T90), identify spectral bands that show significant differences to tentatively associate with molecular fingerprints (biomarkers). (**D.2.**) Refer to known bibliography concerning spectral absorbance bands, the before mentioned information, and make use of peak spectral bands that show significant differences between groups to tentatively associate with molecular fingerprints (biomarkers). (**E**) Raw data is pre-processed and information is fed to different workflows. First, clinical variables are evaluated using a vast toolset including statistical inference, nomograms, decision trees, and others. (**E.1.**) The main workflow dedicated to a comprehensive machine learning process allows for the automatic results and analysis of information from (**E.2.**) PCAs, HCAs, algorithm test and scores, and many more.*

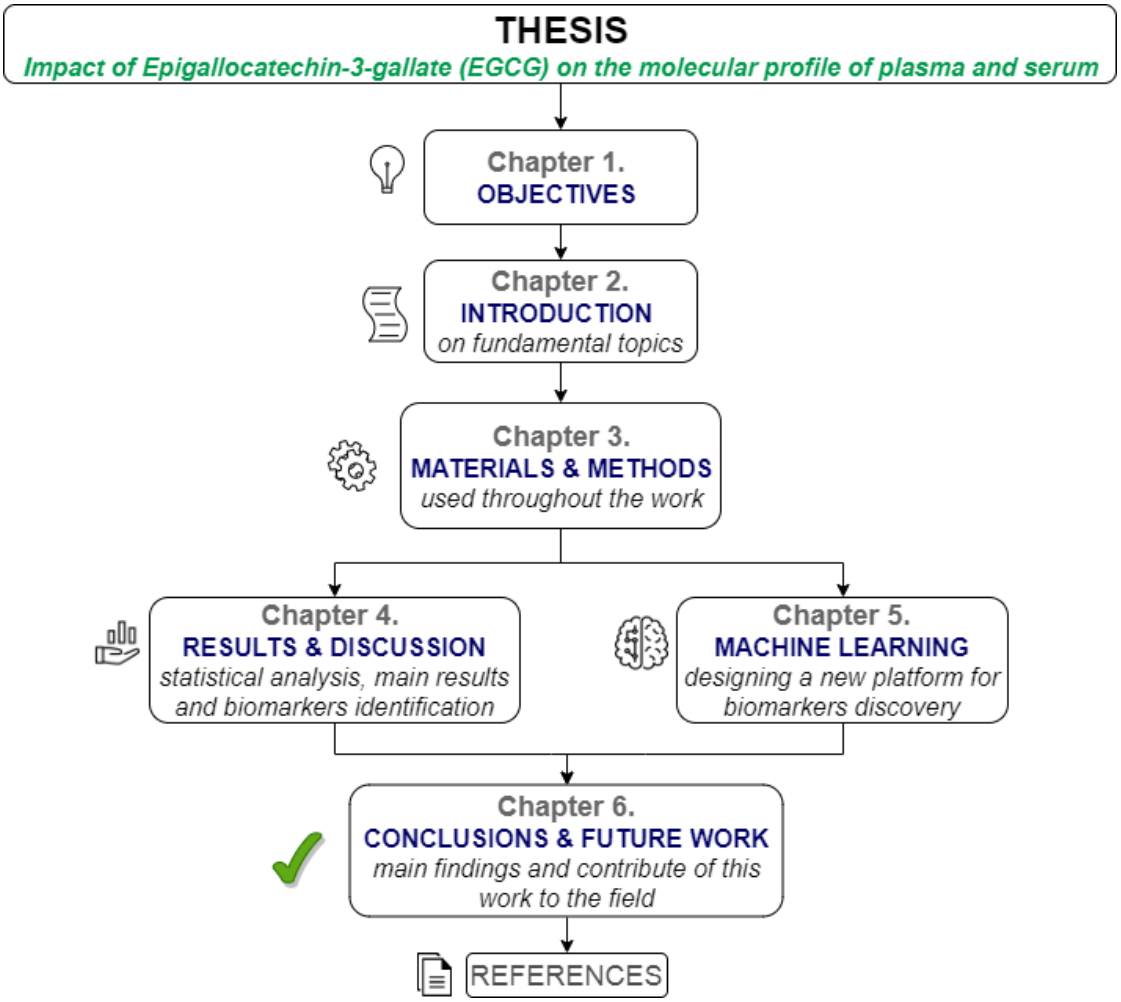


Figure 1.1.2. Block diagram flow chart of thesis outline.

Chapter 2: Introduction

2.1. Tea and epigallocatechin gallate

An historical background

The history of tea is not an easy one to tell. In fact, if we were to study its origins, instead of being able to tell a detailed and concise history, since its discovery and subsequent varieties to its expansion throughout the globe, we would soon find ourselves in a convoluted maze with many different versions, timelines and their very own legends, in which all claim to be the one true story, as often happens with things that date back millennia into our history as a civilization.

Having said that, there seems to be a somewhat agreeable consensus that the history of tea [2], [3], although being often attributed as a quintessential drink of the British [4], has a far more reaching history, being widely accepted that its roots are tied with China [5], a staple of ancient Chinese culture and medicine. It is thought that tea came to be around 2737 BC, during the reign of Emperor Shen-Nung (also known as Shennong, Wugushen, Shen Nong Ben Cao Jing and others). Considered by many to be a mythical sage, a deity of sorts and the father of Chinese agriculture and medicine no less, being attributed the authoring of the earliest surviving work in Chinese pharmacopeia entitled ***Shen-nung pen ts'ao ching (Divine Farmer's Materia Medica)*** [6], which contains 365 medicines that were derived from plants, animals and minerals.

It is believed that much like the story of Isaac Newton, with the apple and the discovery of the law of universal gravitation in the 17th-century, it was also by a similar fashion that green tea came to be, albeit with a more mystical twist to it. Shen Nung, when resting during one of his travels with his convoy, had a few tea leaves fall into his cup of hot water, from a burning tea twig that was in the vicinity. The story goes that the water turned dark but that it went unnoticed by the emperor, after which time he drank it and finding it to be quite refreshing, henceforth became a requested beverage to be prepared by his troops. Although there are other stories and legends that attribute the discovery of the *Camellia sinensis* [7] plant to the Shang Dynasty [8] as a medicated drink or to the founder of Chan Buddhism [9], [10], *Bodhidharma* [11] and even in earlier years, as far as 3000 BC (where it is thought people would use it for chewing and eaten for recreation, much like it was firstly with the discovery of coffee beans in Ethiopia [12]). Regardless, the story of Emperor Shen Nung still remains as the most disseminated story, made famous through the first recorded work in the world, regarding green tea, in Lu Yu's ***Cha Jing* (*The Classic of Tea)*** [13] . In it, it is possible to find a detailed account of the various methods of brewing tea (at the time, an unoxidized version of tea). The tea leaves, crushed immediately after steaming, would then be compressed into layers.

Lu Yu, abandoned as a baby at a temple and adopted by a monk, lived in China's Tang dynasty between 733 and 804 AD, a highly educated scholar and traveler of all of China, learning all there was about the art of tea and sharing his knowledge with people, is said to have lived a peaceful life, despite

his high status and fame within the country. Lu Yu is often regarded by many, including current day tea masters, as the one true sage of tea and even "The God of tea", not because he was the first one to discover it, as he was not, but because of his detailed account of tea in his book (that spanned ten chapters), the perfecting of the craft and his love for the beverage, having often proclaimed his love for tea in his words: "***Its liquor is like the sweetest dew from Heaven***". Although it was during Tang dynasty (that span from the 7th to the 10t century) [14], that tea drinking became the norm and tea ceremonies an integral part of society. There are of course, many other major make or break moments in tea history, namely throughout the 3rd and 6th century Wei Jins' Northern and Southern dynasties (where by the end of the 3rd century, tea had already become China's preferred beverage) and the Ming dynasty (14th – 17th century).

China would open this commodity for trade in the 8th century, trading it with Tibet, Turkey, nomad tribes in the Indian Himalayas and making use of the Silk Road into India as well. Europe would have to wait for the popular beverage until the 16th century whereas the British endured an additional century. As contrary to popular belief, tea was not introduced into Europe by the British, but through Holland and it was the Portuguese princess Catherine of Braganza (*Catarina de Bragança*) [15], that would ultimately introduce tea in England. Later on, in 1840, Anna Maria Russel, the seventh Duchess of Bedford [16], would be credited for what is today known as the afternoon tea, first started as a way to satiate hunger and now a wholesome ritual for the many.

It is important to note something about tea itself, which is that for centuries all tea was green tea. Green tea is, to put it plainly, just the leaves of the *Camellia Sinensis* plant placed in hot water. Unlike today, the leaves would not undergo the oxidation process, giving therefore the green coloring to the water. It was not until the globetrotting travels of the delicate leaves from China to Europe, that damage in the form of oxidation became apparent and that profit-conscious tea producers and distributers searched for ways to maintain the freshness and potency of the leaves. This gave rise to a new process, whereas oxidation of the leaves would occur in a natural process before drying them, giving a darker color to the brew. This would become later know as *black tea*.

In fact, all teas be it green, black, white, oolong, *pu-erh* (or *pu'er*) and yellow tea, stem from the same *Camellia Sinensis* plant, native to China. Tea has become a staple and appreciated in many countries around the world [17] in all its varieties and serving methods, from simple tea, to iced tea, sweet tea and in delicacies like tea cakes. Albeit being known for its medicinal purposes for centuries, it was not until recently in our history that the real medicinal value of tea has been minutiously researched for its potential for antiviral, antioxidant, anticancer properties and treatments (among many others), which will be discussed further along this work. For more about the intricate history of tea we divert the reader's attention to further bibliography [18]–[22].

Tea: general information

The benefits of tea consumption, for all its varieties and methods of processing has, long since its inception in world culture, been documented and its impact in human health proved many times over and still continues to be so to this day. In fact, a simple search for the term "*green tea*" in one public database such as PubMed [23] for example, reveals, at the time of this writing, 7988 articles, in which the various properties long thought to be had by green tea for millennia are brought to light. It is not the intent of this work to extensively review green tea, a work already done and with great success by others. Instead we present a quick summary of the characteristics of tea, followed by a special focus on green tea and some of the research done into its benefits across the various scientific research fields.

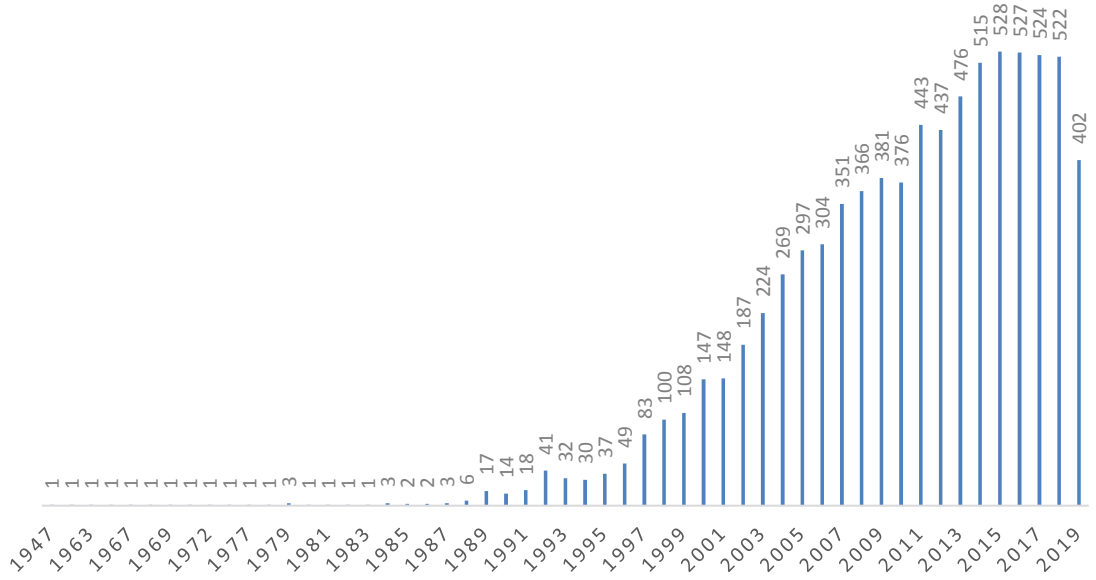


Figure 2.1.1. Results for the number of published articles with green tea in either title and/or abstract". In the x-axis the year of publication and in the y-axis the count number for that year. In 2019, as of the 8th of August of 2019, there were 402 articles published in the PubMed database.

Although the love of coffee by many, being preferred in Europe and in the Americas, it is indeed tea the most consumed beverage in the world after water [24], which can be partly explained by the fact that, since its discovery in China over 5000 years ago, people soon became aware that the consumption of tea promoted health and even seemed to provide respite to human ailing and even prevent diseases. Indeed, a study revealed that tea rehydrates as well as water, due to its content on components such as flavonoids (which account for 30% of the dry weight of a leaf). It also may promote health, being suggested that the regular consumption of 2 to 3 cups per day could provide protection against certain illnesses and types of cancer [25]. These will be discussed further along. Today, there are 195 countries in the world [23], and out of these ones, the majority will have been introduced at one time or another to the beverage. Depending on the source and the year chosen, things can and do vary somewhat in regards of the biggest consumption countries (per quantity or *per capita*, the ones that import/export the largest quantities of tea), but some of the main contenders are always the same: China, Russia, Turkey, with this last one being as of 2016 the leading force in annual *per*

capita worldwide tea consumption [26]. For more information on statistical data, patterns of consumption and how the tea industry shaped the word we refer to [27]–[29]. Unbeknownst to many, and as briefly mentioned before, although today there are hundreds of cultivated varieties and hybrid plants, that have evolved over time, all tea originates from the same plant species, *Camellia Sinensis* [7]. There are two main varieties of the tea plant, namely *Camellia Sinensis Sinensis* and *Camelia Sinensis Assamica*. The first being a smaller-leafed variety typically used to make green and white teas, found in drier, colder climates and the latter being a larger-leafed variety, used to produce strong black teas and is usually found in warm and moist climates such as India and sub-tropical forests.

The way the different teas come to life, has much to do with the way they are processed. For example, with green tea, the leaves harvested from the *Camellia Sinensis* are quickly heated and dried to prevent too much oxidation from taking place. However, if this oxidation is allowed to take place, the leaves will change from green to colors that vary between amber to red or dark brown, giving rise to the famous black tea, where the leaves are fully oxidized before they are heat-processed and dried. At first, tea was thought to also contain the world's most consumed psychoactive drug, a central nervous system (CNS) stimulant of the methylxanthine class [30]: caffeine. This molecule [31] was discovered in 1819 by a German chemist, Friedlieb Ferdinand Runge, while he was studying coffee and its contents, promptly naming the molecule caffeine. However, later tests conducted by chemists on tea revealed that it was not caffeine, but a somewhat similar molecule that was found in tea: theine. It is also important to not confuse the similar sounding terms of theine and theanine. While both are found in tea, theine is a methylxanthine, acting like a stimulant, whereas theanine is an amino acid, which reduces stress and relaxes the body [32]–[35].

While it is common to find in literature the term caffeine attributed to both coffee and tea alike, as previously mentioned, molecules that pertain to both are not quite the same. However, and for the sake of simplicity, from here on out, we will use the term caffeine even when referring to teas' theine. Throughout numerous studies in the last decades, it has been proved that the caffeine biosynthesis has evolved independently both in *Coffea* (plant) and in *Camellia Sinensis* (tea plant) [36]. Regarding the amount of caffeine, there is significantly less in the average cup of tea, when compared to coffee, especially when considering the teas that are brewed at shorter times and cooler temperatures, like white and green teas. However, the high levels of antioxidants present in tea make it so that the absorption of caffeine is done at a slower rate, contributing to a smoother increase of the chemical in our bodies, proving both an increased period of alertness with no withdrawal and crash symptoms in relation to caffeine levels from coffee. If one would choose to drink tea without any amounts of caffeine, then the right choice would be to switch to an herbal infusion, such as *Chamomile*, as these are made from botanicals and not related to *Camellia Sinensis* and therefore are naturally caffeine free. For more information on this topic we recommend the bibliography found in [37]–[43].

Green tea

Green tea, with the scientific classification and binomial name of *Camellia sinensis* (L.) Kuntze [44], accounts for around 20% of the total global tea production. It belongs to the kingdom *Plantae*, the order *Ericales* of the *Theaceae* family. Its genus *Camellia* is comprised of more than 325 different species, although only two of them are commercially viable for producing tea: *Camellia Sinensis Sinensis* and *Camellia Sinensis Assamica*. Green tea, like the other varieties, is an aromatic stimulant, comprised of various polyphenols, essential oils and caffeine, with a varying percentage according to the steep time, method of preparation, variety, '*terroir*' (i.e., environment the tea is grown in) and others, containing a greater percentage of caffeine than coffee per dry leaf weight. There are many different ways to prepare it but the most traditional ones are by the Chinese, which involves harvesting the tea leaves and then quickly heating them by pan firing or by steaming, the Japanese preferred way. After pan firing or steaming, the leaves are then dried in order to prevent oxidation from turning the leaves into a darker color, allowing the leaves to keep its characteristic green coloring and freshly picked distinct flavor. There are dozens of different kinds of green tea [45]. Below we present a small list with some of the most important and well recognized varieties.

Table 2.1.1. Varieties of tea per countries (China, Japan and Korea).

Country of origin	Denomination	Brief description	References
China	*Longjing* (Dragon well)	Tea is pan-fired and has a toasty taste and a smooth, sword-like shape. A classic green tea flavoring, unreproducible by any other tea producing region.	[46]
China	*Gunpowder*	Each leaf is rolled into a small round pellet, giving it the appearance of gunpowder, hence the name. It is fired in a perforated metal tumbler that tosses the leaves around according to a specific pattern.	[46]
Japan	*Sencha*	Makes up for more than 80% of tea produced in Japan, being the most popular tea drunk in households. Steamed tea leaves, rolled into long strands.	[46], [47]
Japan	*Matcha*	Tea leaves are ground into a fine powder, instead of shaped and rolled. Matcha is famously served in Japanese tea ceremonies and is used extensively as a popular ingredient in cooking.	[46], [47]
Korea	*Jeungje-cha*	Green tea prepared with steamed tea leaves, vivid in color. Used in temple cuisine.	[48]
Korea	*Jungno-cha*	One of the most famous Korean green teas. It is a roasted variety of tea, whose tea leaves grow among the bamboo in Gimhae, Hadong and Jinju in the South Gyeongsang province.	[48]

Green tea has many different conditions of growth, methods of cultivation, preparation and has a variety of uses, as we will soon dwell into. The immense capability of green tea to differentiate itself into its many varieties and its use in the treatment of the various ailments is explained, not only by the careful and precise tweaking of the methods one can prepare the serving of the beverage, but also for its uncanny complexity of compounds and chemical composition. Different bibliography tends to show

slightly different values and compounds, due to different methodologies applied, equipment used and even the evolution of research and technology in this field. It is also important to bring attention to the fact that the chemical composition of tea may vary substantially in regards to its genetic strain, soil properties, climate conditions, season in which the leaves are picked and processed, the way they are stored, etc. With this in mind, we present data in Table 2.1.2 regarding the general green teas' chemical composition. For the curious, the chemical composition of green tea (*Camellia sinensis*) infusions, commercialized in Portugal can be seen in [49].

Table 2.1.2. Chemical composition of green tea [50]–[52].

Compound	Brief description
Catechins or Flavan-3-ols	About 30% of the dry leaf weight.
Tannins	Polyphenolic biomolecules that give tea a slight acidity or bitterness of taste (astringency). They play an important role in the ripening and aging of the plant.
Theaflavin	Dimer molecules of catechins, that are responsible for producing red and brown colors and granting astringency to the tea.
Thearubigins	Most abundant polymeric polyphenols formed during the oxidation process, accounting for up to 60% of compounds in oxidized tea and red coloring. Increasing amounts of concentration will cause a change in color from amber to dark brown.
Vitamins	Vitamins B2, C, E, folic acid and β-carotene.
Saponin	0.1% saponins, which give it a strong bitterness and astringency. Saponins have anti-fungal, anti-inflammatory and anti-allergenic properties.
γ-aminobutyric acid (GABA)	GABA is formed in tea when the raw leaves are left without oxygen.
Minerals	Ca, K, Mg and smaller quantities of Zn, Cu and Mn.
Caffeine	10 mg of caffeine per 100 mg of leaves.
Carbohydrates	40% of carbohydrates, a third being cellulosic fiber.
Lipids	4% of the total green tea leaves.

Around 4000-4000 years ago, the Chinese became aware of the fact that tea could promote human health and even prevent diseases, like described in *Shen Nong's Tea Classic* [13]. Since then it has been widely studied across the globe and are nowadays accompanied by a growing body of evidence regarding tea health-promoting effects and effectiveness in human diseases, e.g., cardiovascular and neurodegenerative and for displaying antioxidant, anti-inflammatory, antiviral, antimutagenic, anticarcinogenic, antibacterial and antifungal properties [51]–[68]. Green tea consumption has also been said to alleviate depressive symptoms in the elderly population [69]. It is important to note though that problems can arise with over consumption of green tea. In fact, excess consumption of green tea can cause several health complications. One of the catechin-based flavonoids in green tea leaves, namely the epigallocatechin-3-gallate (EGCG), is cytotoxic. At high consumption levels it can exert acute levels of cytotoxicity and hepatotoxicity in liver cells, among other major metabolic organs in the human body. High concentration of catechins can also provoke a deficiency in iron bioavailability and cause goiters. Other problems include sleep disorders due to increased levels of caffeine, headaches, epigastric pain and tachycardia. Some studies also suggest the possibility that green tea

may interact with some supplements, e.g., in combination with stimulant drugs, green tea could increase blood pressure. It can also interfere with medications, and drinking green tea should be avoided when in conjunction with medication such as blood thinners like Warfarin: Same with aspirin, as the vitamin K existent in green tea also acts as a natural blood thinner, therefore further reducing the clotting effectiveness of platelets [70]–[72]. Excessive consumption can also lead to autoxidative reactions and result in ROS (Reactive oxygen species) production [73]. From these studies came the decision by nutritionists of recommending a regular and controlled consumption of two to three cups of green tea per day, in order to reap the beneficial health effects of the beverage, while mitigating the dangers that are associated with overconsumption [73]–[80]. For more information about the possible harmful effects the article from the European Food Safety Authority may be consulted [79], the Toxicology Data Network of the USA National Institute of Health (TOXNET) [81] clinical trials website of the USA National Institute of Health (NIH) [82].

Epigallocatechin Gallate

The study of green tea has intensified for the past 20-30 years, much due to its health benefits, some of them know for millennia. The tea constituents have been demonstrated to show various biological and pharmacological properties that grant protective effects such as being antioxidant, anticarcinogenic, antiallergic, antiviral, hepatoprotective, prevent tooth decay, diabetes, etc. These properties have been associated to polyphenols such as catechins. The major green tea catechins are (–)-epicatechin (EC), (–)-epigallocatechin (EGC), (–)-epicatechin-3-gallate (ECG) and (–)-epigallocatechin-3-gallate (EGCG), with the latter being the most abundant in green tea leaves and the most bioactive catechin. A comparative study between the antioxidant properties of the main catechins and other oxidants can be found in [83]. The antioxidant properties of polyphenols are of great importance due to their redox properties, which allows them to act as reducing agents, hydrogen donors and singlet oxygen quenchers. EGCG can only be found in the tea plant and as such it is regarded as the main characterizing constituent of green tea.

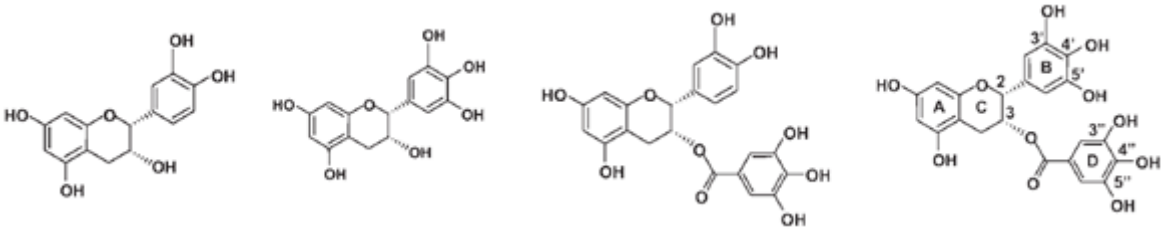


Figure 2.1.2. Chemical structure of the major catechins in green tea. From left to right: epicatechin (EC), epigallocatechin (EGC), epicatechin-3-gallate (ECG) and epigallocatechin-3-gallate (EGCG).

Catechins, generally stable in solutions at pH values between 4 and 6, comprise two benzene rings (A and B), as pointed out in Figure 2.1.3, and a dihydropyran heterocycle, the C ring, with a hydroxyl group on carbon 3. As for the tea catechins, they are characterized by a dihydroxyl or trihydroxyl substitutions on the B ring and the *m*-5,7-dihydroxyl substitutions of the A ring [64]. Catechins are known to be bound and transported by the human serum albumin [84].

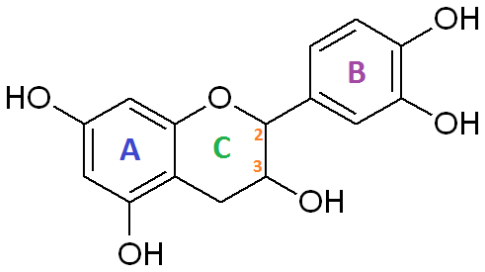


Figure 2.1.3. Chemical structure of a catechin.

In order for these catechins to have an effective role in the human body they need to be bioavailable after they are consumed. Bioavailability refers to the extent and the rate at which an active compound enters the systemic circulation and gains access to the site of action [85]. Both extent and rate ultimately depend on the properties of the compound, dose levels, the frequency and the type of administration. Bioavailability has been studied when tea is consumed with other food products such as milk, the English preferred combination since the 17th century. A time when it was poured before the tea, as a way to avoid cracking of the poor quality china cup teas [86], due to the high temperatures the beverage is served. The influence of milk added to both green and black tea on the polyphenol absorption has been researched by many groups, some of them observing that the inclusion of milk seemingly reduced the absorption of catechins, pointing the fat content of the milk as the main factor for the poor bioavailability [87], [88]. Others reported that the catechins from green and black tea are rapidly absorbed, with milk not having an effect on bioavailability [89]–[91]. A review of 97 studies regarding the bioavailability and bioefficacy of polyphenols in humans can be found in [92].

EGCG, taken by oral administration, is absorbed and metabolized in the intestine. Like other phenolic compounds, EGCG has a low bioavailability [93] due to its oxidation, metabolism and efflux [94]. EGCGs' bioavailability has been increased through the use of biocompatible EGCG encapsulated nanostructured lipid carriers (NLCE) and other structures [95]. However, there are possible setbacks in using nanotechnology, as the metabolism of engineered nanoparticles can lead to the formation of products that induce hepatotoxicity [96]. As such, the use of engineered nanoparticles is still quite limited and a relatively new technology. Others, have reviewed different factors [97] that can enhance plasma levels of ECG (albumin, vitamin C, etc.) or that diminish it (air content oxidation, sulfation [98], gastrointestinal inactivation, etc.). Once the catechins enter the organism, they undergo metabolic processing in the liver, small intestine and colon.

In humans, plasma bioavailability of green tea catechins is quite low [99], [100]. The half-lives of flavanols are around 2-3h in plasma, except for EGCG, which is thought to be slower due to high biliary excretion and interaction with plasma proteins [101]. The native forms of ECG, EGCG and the corresponding metabolites of EGC and EC can be found and its concentration measured in blood plasma, but not in urine. In fact, no forms of ECG and EGCG are detected at all in urine, only the before mentioned metabolites of EGC and EC. This inability to detect EGCG in urine, despite being detected in plasma, has been studied, with some researchers hypothesizing that it could be that the

kidneys are unable to clear EGCG from the bloodstream, although the methods are still not quite understood [102]–[104]. EGCG is eliminated mostly by biliary excretion [105] and the colon, with the majority of EGCG ingested not getting into the blood stream. ECG and EGCG is not detected in urine but the same is not true for EC and EGC with amounts up to 90% being excreted after 8h, followed by a continuous decrease in catechin levels. 24h after consumption, the levels of catechins have been reported to not be detectable [106]. Other authors describe the urinary recovery to be between 0.5 to 6% for some tea catechins [107]. The authors in [60] reviewed several articles and provide useful tables with the amount of EGCG in blood plasma after a single dose (that goes as low as 50 mg to as high as 1600 mg of EGCG), as well as the amount of EGC in a 24h urine collection.

2.2. Biomarkers

Biomarkers are basically measures of a certain biological state [108]. There are however slight variations of its definition according to different authors. Based on BEST (Biomarkers, Endpoints and other Tools) [109], created in 2015, between the Food and Drug Administration (FDA) and the National Institutes of Health (NIH). A biomarker is "*a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions*".

In a PubMed search, the first result of a search for '*biomarkers*' as *title/abstract* comes up in 1973. It was only in 1987 that the number of publications broke the 2 digits, with 19 publications. Today, in 2019, and at the time of this writing, the number is already up to 17461. This makes up for a total of 157507 published articles in the last 46 years, with approximately a third of these (53437) having been published in the last 5 years and concerning biomarkers found in human studies.

The interest in this field is not showing any signs of slowing down in the near future either with many published articles about their potential uses [110], [111]. Inclusively, a recent study in the metabolic profile of 44,168 individuals, published in Nature Communications (August of 2019), has identified a group of 14 metabolites, where increased values in these are associated with a 2.73 times higher mortality risk. These results may ultimately allow pave the way for a more precise prediction of risk of death that the traditional health risk factors [112]. Biomarkers research is a vast field of biomedical research and is characterized by a clear separation of the different categories of biomarkers as can be seen below in Figure 2.2.1. In the figure we have highlighted in light green the category at which we have classified the biomarkers sought out and studied in this work.

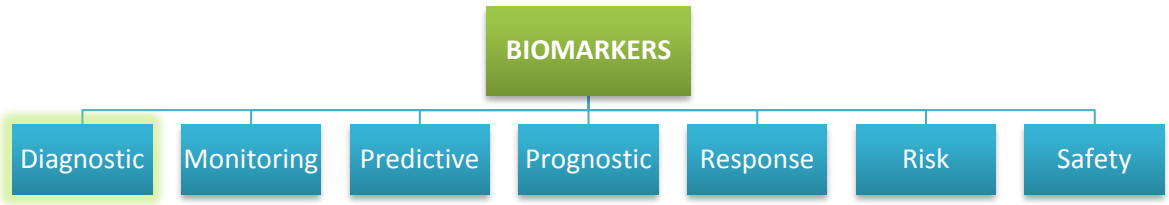


Figure 2.2.1. The different categories of biomarkers.

The principles of biomarkers have been applied to detection, screening, diagnosis, treatment and monitoring of many diseases including many types of cancer. According to an article published in 2001 [1], there are five phases of biomarker development. They can be summed-up as shown in Figure 2.2.2. It is important to mention that, when attempting to develop a biomarker, either to check for a particular disease or human condition, in clinical trials [108] or any other of the five phases, one should always strive for the ideal biomarker attributes: safe and measurable; easily modifiable, cost-efficient and consistent (over gender, ethnic groups, etc.).

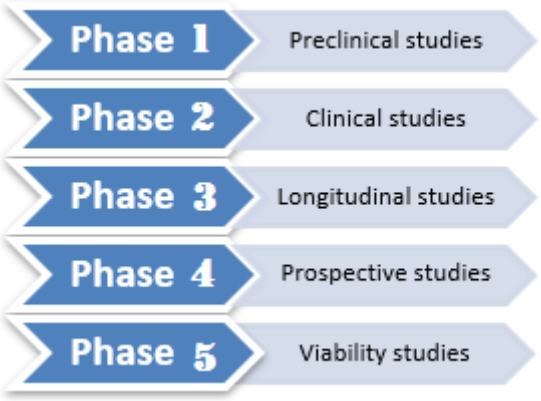


Figure 2.2.2. The five phases of biomarker development for the study of cancer.

2.3. Plasma and Serum

Currently, the large majority of search for biomarkers focus biofluids as plasma and serum, most due to their easy accessibility on biobanks [1] but also as presenting a high metabolic information. For example, it is quite easy to find articles that make use of MIR and FTIR spectroscopy to search for biomarkers [113]–[115] and more specifically that screens plasma [116], [117] and serum [118].

Both serum and plasma come from the liquid portion of the blood that remains once the cells have been removed. Serum exhibits a light-yellow color and is the remaining liquid after the blood has clotted, in which fibrinogen (a clotting factor) can be found. Plasma, exhibiting a somewhat clear appearance, is what remains when clotting is prevented with the use of anticoagulants. Plasma makes up for about 55% of the overall blood volume. It is the liquid portion of the blood with around 90% of it being water.

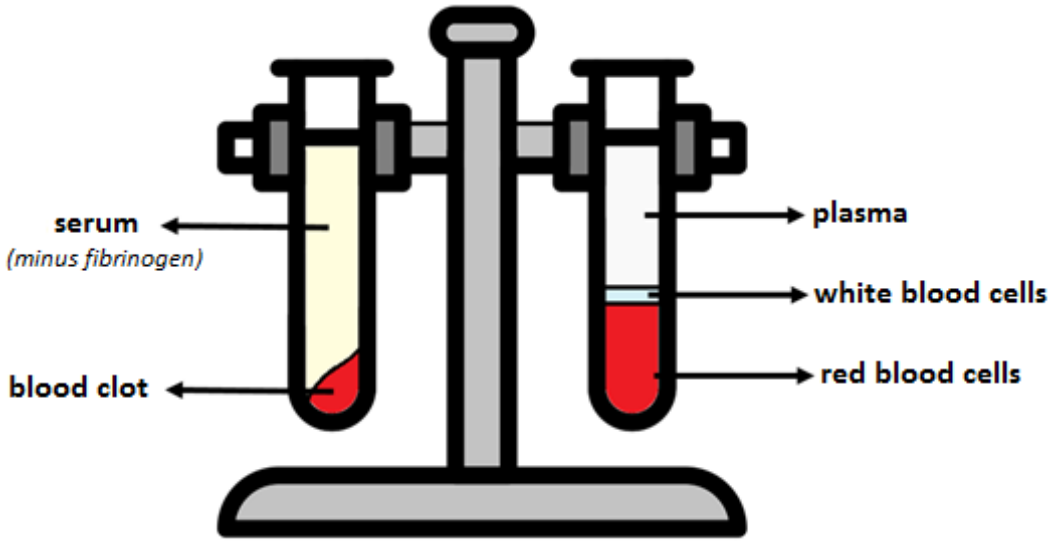


Figure 2.3.1. Main compounds of plasma and serum.

Human plasma [119]–[122] and serum [123]–[125] are widely used in biological and clinical studies. Some studies pointed out that the most informative biofluid was plasma [126], while others singled out serum [127]. One study refers 2D-FTIR spectrometry as a useful tool to analyze plasma and to correlate with molecular species as lactate and proteins [128], where a group of swimmers was subject to intense physical exercise, after which blood samples were taken and it was observed exercise-induced changes in plasma concentration and plasma spectra absorptions. One other research team has observed that there was a good reproducibility in both biofluids regarding repeated measurements of metabolite concentrations, though the reproducibility was slightly higher for plasma than for serum (p = 0.01, paired t-test), with mean correlation coefficients (r^2) of all 122 metabolites of 0.83 for plasma versus 0.80 for serum. However, the authors of the study also observed that for 104 out of the 122 metabolites (85%), registered a higher metabolite concentration in serum [129]. As such it is thought that serum provides more sensitive results, therefore more useful when searching for biomarkers. For more information about the chemical composition, isolation procedures, advantages of using either plasma or serum, etc., it is recommended the following bibliography found throughout [130]–[137].

2.4. Omics Science

Traditional technologies use a reductionist approach to solve problems, but this is rarely sufficient to solve highly complex biological problems. Instead, what is needed is to see the bigger picture, using a holistic and integrative approach to better figure out the whole process. This is brought by systems biology and is the main core of what is known as the Omics technologies, valuable tools that allow for comprehensive analysis, both qualitatively and quantitatively of a target biological process. Today, automated DNA sequencers, microarrays, mass spectrometry and infrared spectroscopy technologies allow for a global transcriptional profiling and large-scale proteomic and metabolomic analysis.

There is a growing body of works regarding omics that use plasma and serum. From the quantitative characterization of serum, e.g. metabolomes of children that link to molecular changes [138], multi-omics analysis in metastatic melanoma [139], proteogenomics for the comprehensive analysis of cellular and antibody repertoires [140], profiling leptospirosis patients to investigate proteomic alterations [141], identification of biomarkers associated with migraine [142]. Other omic studies use plasma as a diagnostic fluid, e.g. for male reproductive system disorders [143], to predict insulin sensitivity in obese, nondiabetic individuals [144] and use it for sport and exercise science [145].

Since their inception, omics have taken over the world by storm as major tools of biology and medicine [146], [147], [148]. In the US, the National Institutes of Health Metabolics Common Fund has even funded a *metabolics database* [149]. In omics, a large number of molecules are comprehensively characterized and quantified. They are then separated by groups according to the structural or functional similarities they may exhibit. There are many, but the most important and largest categories are represented in Figure 2.4.1.

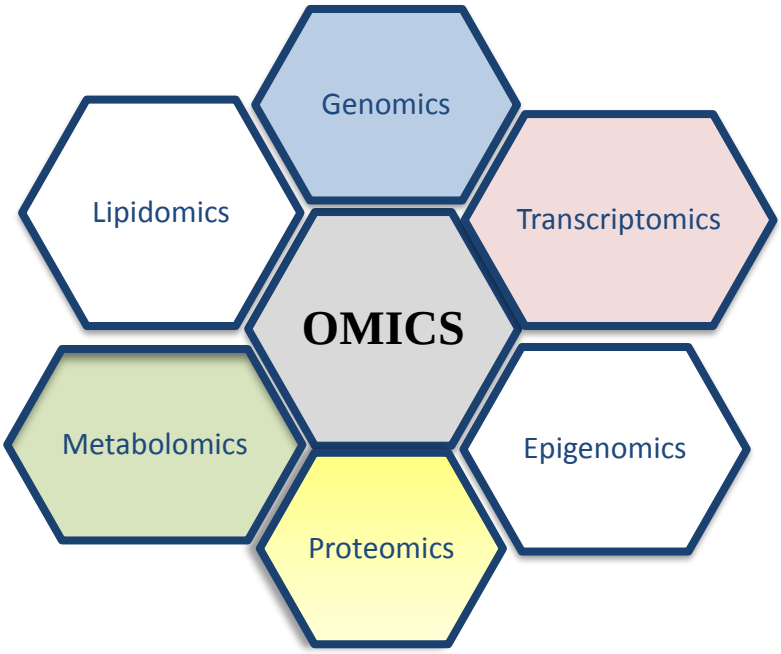


Figure 2.4.1. Overview of the major omics fields.

In Figure 2.4.1, the largest categories in omics are represented, from which the color highlighted, the four major omics fields: genomics, transcriptomics, proteomics and metabolomics. A more detailed description of systems biology, advantages and disadvantages of omics can be found in [150], but there is one disadvantage that is the focus of the next paragraph: speed, or lack thereof.

With the rapid advances in technology and informatics, as well as the adoption of high-throughput omic approaches to analyze biological samples such as genomics, transcriptomics, proteomics and metabolomics, a critical shift has been seen in the study of biomedical sciences [151]. This is giving rise to inter-disciplinary data integration strategies and teams, in an effort to better understand

biological systems and eventually develop more successful and precise medicine. However, this is also creating a problem, though not new, of unseen proportions. Today, the analysis of biological data through omics can easily generate up to peta-bye sized data files, making the integration and translation of these multi-dimensional omics data, into biologically meaningful data, a complex and challenging task [152]. The future is looking bright though, with the advent, sooner than most think, of quantum computing. As the time of the writing of this text, Google is said to be on the verge of quantum supremacy, a term that defines the solving a complex problem that would otherwise be impossible by normal computation. To understand the relevance of this future capability to the omics field, here is a simple example. A computation done on the most advanced supercomputer, Summit, would take 10,000 years versus the alleged 3 minutes and 20 seconds of Google's new Sycamore quantum computer. Alleged as the mysterious article was promptly removed from NASA's website minutes later, though the internet being as it is, has already made numerous copies of it [153].

Regarding metabolomics (the omics field most prominent perhaps in this work) it can be described as the "*nonbiased identification and quantification of all metabolites in a biological system*" [154]. The interactions between the metabolites and their surrounding biological system are known as the metabolome. Metabolomics importance is undeniable as they have a close relationship with the phenotype. This makes the study of metabolomics of paramount importance to preventive healthcare, the pharmaceutical industry and represent a gateway for biomarker discovery [155] and drug safety screens [156]. Regarding the use of this technology when studying green tea, there have been several and diverse studies, though often only concerning discrimination of tea varieties [157], the metabolomics of diverse green tea *cultivars* [158], metabolic and functional roles of flavonoids in light-sensitive tea leaves [159], authentication of geographic origins [160].

In revised bibliography it was observed to be rather easy to find green tea and metabolomics as the main focus of each paper, the techniques rarely touched upon spectroscopy, usually falling under techniques such as mass spectrometry or liquid chromatography [161]. The technique used in this work, FTIR spectroscopy, was usually found having green tea as the object of study, but not regarding its metabolomics per say, focusing instead on the classification of the effects of polyphenolic compounds on cancer cells [162], or the effects of green tea catechins on various human conditions and diseases [163].

To the best of knowledge of this working group, never before have the changes in metabolomics of human plasma and serum, after the ingestion of green tea EGCG extract, have been attempted to characterize, quantify and group, using FTIR spectroscopy as its main technique and several supervised and unsupervised learning methods. There are however publications which put emphasis on more traditional metabolomics methods like using HPLC-FTIR (High Performance Liquid Chromatography – FTIR), to determine, qualitatively, catechins and methyl xanthines present in green tea extracts [164]. One FTIR study focused on green tea leaves and their diseases [165], while another on the antimicrobial activities [166]. FTIR spectroscopy has also seen an increase in interest by research groups that are more specialized in using supervised/unsupervised methods, with studies

using Artificial Neural Networks (ANNs) to assess wine quality [167], identification of artificial sweeteners [168], studying the spread of the malaria vector [169], creation of machine learning applications for classification of chemical spectra [170], to name but a few.

2.5. FTIR Spectroscopy

Theory of Infrared Spectroscopy

In this section it will not described the rich history of spectroscopy nor its road to medical vibrational spectroscopy as reviewed in [171]. There are many different types of radiation and consequently of spectroscopy, being distinguished by the type of radiation involved and the nature of the interaction between radiation and matter, where the main interactions between radiation and matter are based on *absorption* (when matter absorbs the emitted radiation), *emission* (when radiation is emitted by matter) and *reflection* (when radiation emitted by a source is reflected on matter) [172]. The effect of electromagnetic radiation on molecules is different regarding the different spectrum wavelengths (Figure 2.5.1) [173]. In IR radiation, the molecular effects concern the vibrations of the atoms, the base for the IR spectroscopy technique.

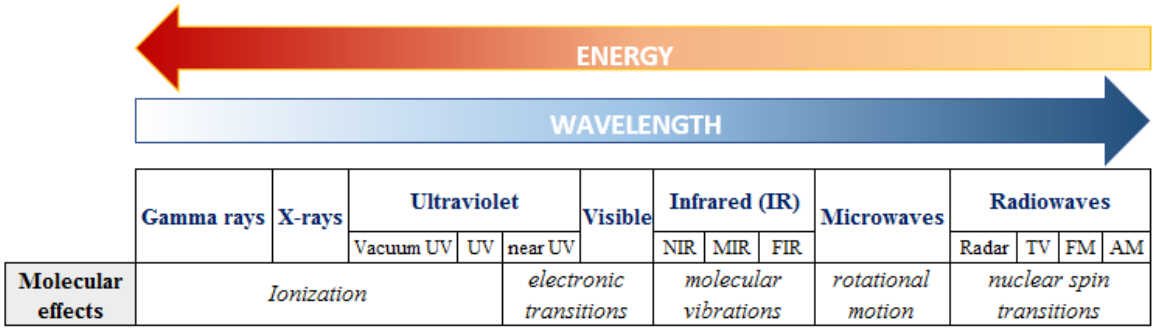


Figure 2.5.1. The electromagnetic radiation effect on molecules.

Infrared radiation is a region of the electromagnetic spectrum that ranges between 780 nm and 1 mm. As observed in Figure 2.5.1, it is usually divided into three main regions: near-IR (NIR), mid-IR (MIR) and far-IR (FIR).The first region, NIR, being the highest portion in the IR spectrum is associated with overtones and combinations of bond vibrations; MIR, the regions which is the focus of this work, is normally used to study fundamental vibrations and associated with the rotational-vibrational structure of the molecules; lastly, FIR is the lower energy region in the spectrum and used for rotational spectroscopy (Table 2.5.1).

The infrared spectrum is usually obtained by the passing of IR radiation through a sample. At this point, it is determined what fraction of the incident radiation is absorbed at a particular energy level. Each peak has a corresponding energy in the absorption spectrum which corresponds to the frequency of vibration of a compound in the sample molecule. This is why IR spectroscopy is so valuable for research, as the analysis of an infrared spectrum can give information regarding the molecular

composition, leading to IR spectroscopy being one of the most important analytical techniques used in chemical and pharmaceutical analysis.

Table 2.5.1. Infrared spectral ranges and characteristics. (adapted from [174], [175]*)*

	INFRARED REGION		
Parameter	NIR	MIR	FIR
Wavelength (μm)	0.78 – 2.5	2.5 – 50	50 – 1000
Wavenumber (cm^{-1})	12.821 – 4000	4000 – 200	200 – 10
Frequency (THz)	120 – 384	12 – 120	0.3 – 0.05
Energy (eV)	0.5 – 1.59	0.05 – 0.5	0.0012 – 0.05

Despite the obvious advantages, no technique is without its flaws. While IR spectroscopy is easy to use and can process samples in the three principle states of matter (solid, liquid and gas), fast, accessible, inexpensive, highly sensitive and produces rich spectra, it does suffer from some setbacks, like not being able to detect some molecules and being sensitive to atmospheric compounds (CO_2 and water), needing extra care in pre-processing and processing of the samples.

Molecular Vibrations

Infrared spectroscopy is a technique based on the vibrations of the atoms in a molecule [176]. Absorptions can be the result of a change in bond length (*stretching*), or the bond length can stay constant, but the bond angles vibrate about their equilibrium values (*bending*). Stretching can be in-phase (*symmetrical stretching*) or out-of-phase (*asymmetrical stretching*). Bending vibrations can include various movements such as scissoring, rocking, wagging and twisting, as can be seen in Figure 2.5.2 [177].

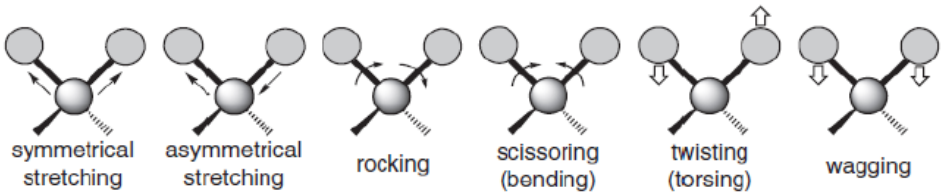


Figure 2.5.2. Possible vibration modes of a molecule. Black arrows concern linear movement in the paper plane, while the white arrows represent movement in and out of the paper plane.

For a molecule to show infrared spectrum, its electric dipole moment necessarily needs to change and therefore heteronuclear diatomic molecules are designated as '*infrared-active*' molecules, since their dipole moment changes (Figure 2.5.3). Opposite to the heteronuclear diatomic molecules we find the '*infrared-inactive*' molecules, such as homonuclear diatomic molecules (e.g., O_2, H_2, N_2) and monoatomic molecules, composed of only one type of atom (e.g., Ar, He, Ne). They are considered to be inactive as their dipole moment remains zero, regardless of how long the bond.

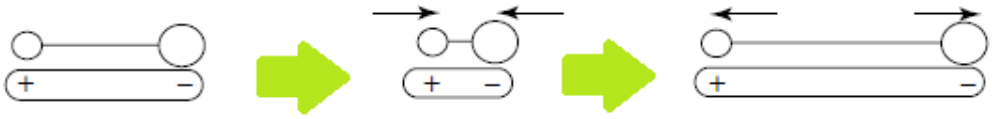


Figure 2.5.3. Change in the dipole moment of a heteronuclear diatomic molecule. (adapted from [176]*)*

The dipole moment of the molecule changes as the bond between it expands and contracts (Figure 2.5.4). The frequency of the stretching vibration is dependent on the mass of atoms and stiffness of the bond. What this means is that lighter atoms vibrate faster than heavier ones. Also, stronger bonds are usually stiffer and compressing or stretching these bonds require therefore more force (e.g., triple bonds are stiffer than double bonds), making stronger bonds to usually vibrate faster than weaker bonds (assuming equal mass). It is important to note that the more complex a molecule is, the more vibrational modes it has (i.e. degrees of vibrational freedom). Also, the larger the change in the dipole moment of a molecule, the more representative the absorption bands will be. For a better visualization on how to properly determine the degrees of freedom for any given molecule, Table 2.5.2 is proposed.

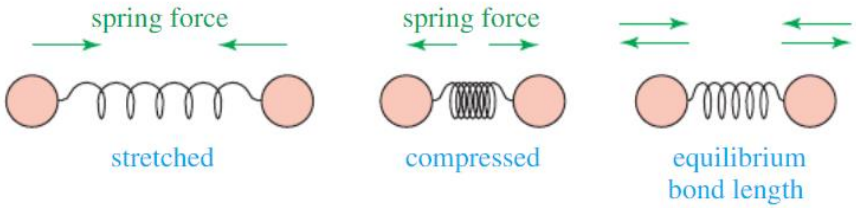


Figure 2.5.4. Specific bonds respond to(absorb) specific frequencies.

In a nonlinear, polyatomic molecule with N atoms, we usually have 3N – 6 fundamental vibrational models. Let us take the example of one of the most know molecules: water. Water has 3 atoms. We then have 3(3) – 6 = 3 fundamental modes as seen in Figure 2.5.5 [173]. This is what makes each molecule unique in their representation on the IR spectrum, as there are no two alike – a molecular fingerprint. This is what allows to determine the composition of a sample through their unique molecular compositions and distinct spectra [174], [176].

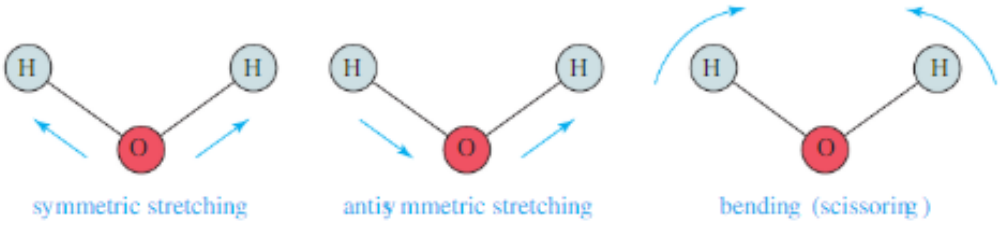


Figure 2.5.5. Fundamental vibrational modes of the water molecule.

Table 2.5.2. Degrees of freedom for polyatomic molecules. (adapted from [176]*)*

Types of degrees of freedom	Linear	Non-linear
Translational	3	3
Rotational	2	3
Vibrational	3N – 5	3N – 6
Total	3N	3N

MIR Spectral Analysis

The mid-infrared (MIR) region is usually plotted between 4000 and 400 cm^{-1} and used to study fundamental vibrations and structure of molecules. This region of the infrared is better suited to analyze biological samples as it is home to stronger, better-defined bands, therefore, it is more sensitive to the molecular composition of the samples and their surrounding environment. MIR spectra can also give us information regarding the conformation changes of biomolecules (e.g., protein folding [178]–[180], nucleic acids [181]–[183]). MIR can also be divided into four regions, which allows to determine the nature of a group frequency.

Table 2.5.3. The four regions of MIR.

	MID INFRARED REGIONS (cm^{-1})			
	X-H stretching	**Triple-bond**	**Double-bond**	**Fingerprint**
Frequency	4000 – 2500	2500 – 2000	2000 – 1500	1500 – 600

There is a great variety of rich bibliography that accentuate the characteristic bands for biological materials [184], may they be proteins [179], peptides [180] and as close to a complete library for spectral interpretation in the mid-IR region [185], medical applications for IR therapy [186] and others [187]. As such, we have only made a short summary of some of the most important bands in the form of Table 2.5.4. In Figure 2.5.6, we have represented a single FTIR spectrum in the mid-IR region for a patient before the ingestion of the EGCG extract.

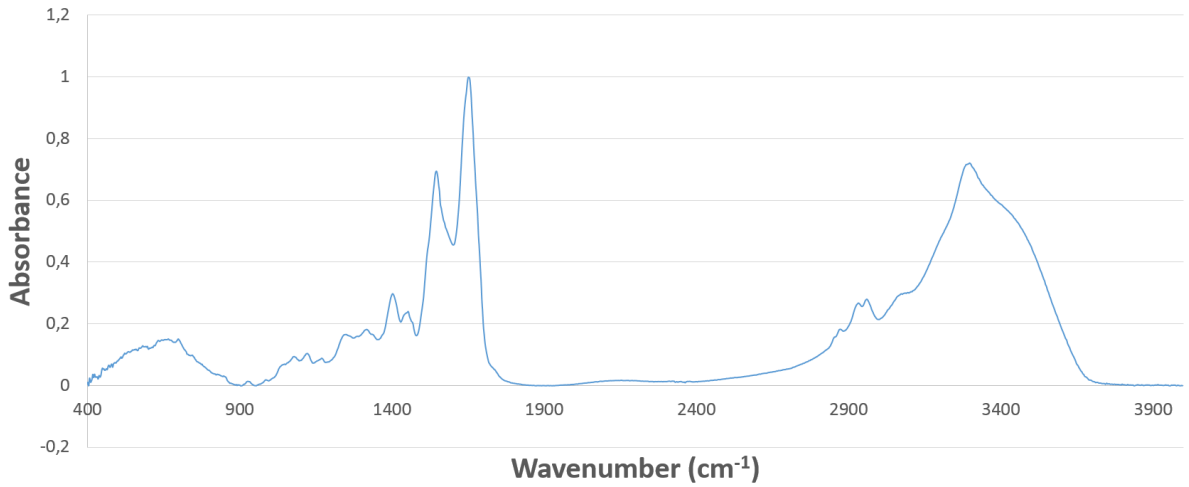


Figure 2.5.6. FTIR spectrum obtained in this work in the mid-IR region of plasma diluted at 1/10 in water of a patient before ingestion of the EGCG extract. The spectrum was pre-processed by atmospheric and baseline correction and normalized to the Amide I peak. The spectrum resulted from 64 coadded scans and at a resolution of $2cm^{-1}$.

To sum-up, spectroscopy in the mid-infrared spectral region provides quantitative chemical, structural and compositional information on the constituent molecules in the solid, liquid and gas phases, allowing for its use in a range of applications from medicine [188], biotechnology [178],

environmental analysis [189], production monitoring [190], materials science [190], etc. In Table 2.5.4, it can be observed some of the important vibrational frequencies in MIR spectroscopy.

Table 2.5.4. Short summary of some of the important vibrational frequencies within the mid-IR region of the electromagnetic spectrum. (adapted from [185], [191], [173]*)*

X – H stretching (where X is C, O or N): 4000-2500 cm^{-1}		
Wavenumber (cm^{-1})	**Functional group**	**Biochemical component**
~3300	N-H	Amide A
~3100	N-H	Amide B
2957	$C\text{-}CH_3$	Lipids
2920	$-(CH_2)_n-$	
2872	$C\text{-}CH_3$	
2851	$-(CH_2)_n-$	
Double bonds stretching: 2000-1500 cm^{-1}		
Wavenumber (cm^{-1})	**Functional group**	**Biochemical component**
~1655	O=C-N-H	Amide I
~1545	O=C-N-H	Amide II
~1740	$-CH_2-COOR$	Phospholipid esters
Fingerprint region (overlapped vibrations): 1500-500 cm^{-1}		
Wavenumber (cm^{-1})	**Functional group**	**Biochemical component**
1400-1200	O=C-N-H, CH_3	Amide III, protein, collagen, DNA, RNA, phospholipid, phosphorylated protein
~1060, 1050, 1015	C-O	DNA and RNA ribose
~1095, ~1084	C-O, C-O-H	DNA, RNA, phospholipid, phosphorylated protein

IR equipment

All IR spectroscopy is performed on instruments called spectrometers. These were initially dispersive instruments that made use of prisms made of materials as sodium chloride. This dispersive element would be found within a monochromator. The first classic dispersive IR spectrometers appeared in the 1940s, although in the 1960s improvements in technology made the use of prisms obsolete in detriment of using diffraction gratings (Figure 2.5.7)

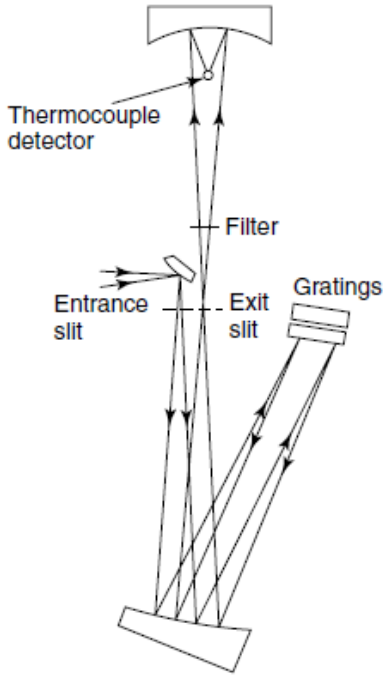


Figure 2.5.7. Schematic diagram of a double beam IR dispersive spectrometer with a grating monochromator as dispersive element.(adapted from[176]*)*

In this type of IR spectrometer, dispersion occurs when energy emitted from the source and focused on the entrance slit, is collimated onto the gratings, at which point the dispersed IR radiation is separated into the different wavelengths of the spectral range and reflected back to the exit slit, beyond which the thermocouple detector can be found. The dispersed spectrum is then scanned across the exit slit by rotating a component within the monochromator. There was a major problem with this setup as the narrow slits at the entrance and exit of the monochromator limited the wavenumber range of the radiation reaching the detector to one resolution width [192]. This limitation would be later overcome through the use of a FTIR spectrometer. FTIR spectroscopy relies on the interference of radiation between two beams to create an interferogram, the resulting signal created by the change of wavelength between the two beams. It relies on a mathematical method known as Fourier-transformation. The basic components of a spectrometer can be seen in Figure 2.5.8.

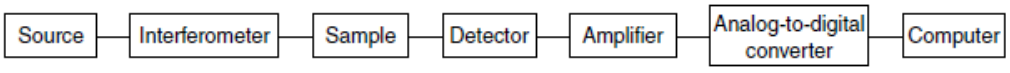


Figure 2.5.8. Basic components of spectrometer. (adapted from [176]*)*

Unlike previous spectrometers, FTIR based ones can measure all wavelengths at the same time. FTIR came to be at the hands of Albert Abraham Michelson, who would end up being the recipient for a Nobel Prize in 1907 for his precise measurements of the wavelengths of light. He would lend his name to the most common interferometer used in FTIR spectroscopy: a Michelson interferometer [176].

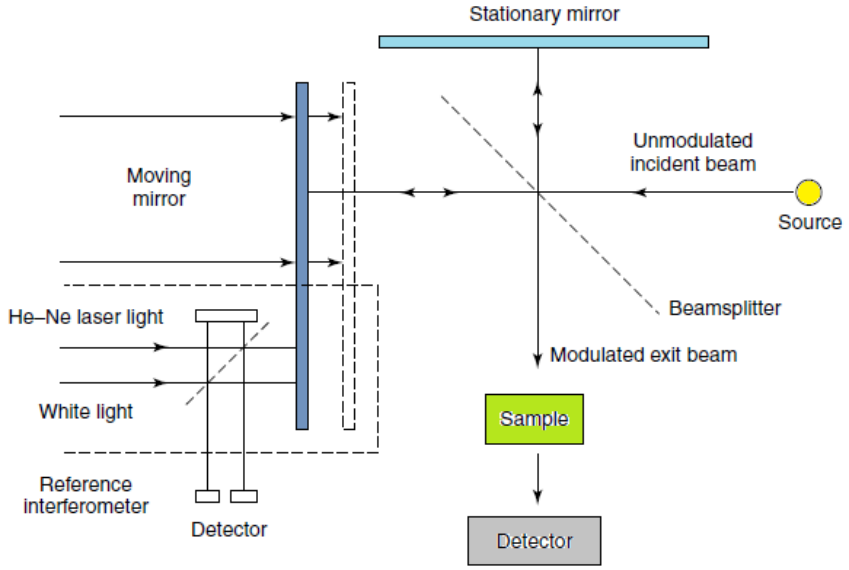


Figure 2.5.9. Schematic diagram of a Michelson interferometer. It consists of two perpendicularly plane mirrors, one of which is able to travel. A semi-reflecting film (beamsplitter) bisects the plane of these two mirrors. The beamsplitter material is chosen according to the region to be examined (potassium bromide or caesium iodide substrates are used for NIR and MIR regions). (adapted from [173]*)*

In the Michelson interferometer, a collimated beam of monochromatic radiation leaves the source and upon reaching an ideal beam splitter, it is divided into two separate beams. 50% of the incident radiation is reflected to one of the mirrors and the other 50% transmitted to the other mirror, which has a mechanism that allows its control and movement by a few millimeters. The two beams then recombine and the resulting single beam is transmitted or reflected off the sample. Given the fact that the two beams cover different distances before being reunited, this results in an interferogram, a signal that is a direct result of the two beams '*interfering*' with each other. This signal is then decoded through Fourier-transform to be able to obtain the final IR spectrum. FTIR spectrometers have significant advantages when compared to older dispersive instruments, as in Table 2.5.5.

Table 2.5.5. Advantages of FTIR spectrometer in relation to dispersive spectrometers.

PARAMETER	DESCRIPTION
Time	Measure many wavelengths at one time, decreasing the spectra acquisition time and resulting in a decrease in noise. This is designated as the *Fellgett* or multiplex advantage.
Sensitivity	Since the total source output is passed through the sample continuously, results in higher gains of energy in the detector, leading to higher signals and improved Signal to Ratio (SNR). This is known as the *Jacquinot's* or throughput advantage. In our work we use 64 coadded scans which result in a noise reduction of 8 (when compared with a one scan spectrum) [176].
Speed	The ability of the mirror to move short distances in a rapid manner, in combination with the SNR improvements by the *Fellgett* and *Jacquinot's* advantages, make it possible to obtain spectra faster (in the order of milliseconds) [192].
Precision	Since a helium-neon laser is used as reference and as such the mirror position is known with high precision, which translates into also being able to determine with precision the position of an infrared band.
Mechanical simplicity	By having only one moving part (mirror), it is a very simple and easy to maintain equipment, requiring no external calibration [172], [173], [176], [193].

The sensitivity, speed of spectra acquisition and the precision of FTIR spectrometers have seen many companies betting on a wide variety of software that make full use of IR radiation for qualitative and quantitative analysis. Today, FTIR has revolutionized infrared spectroscopy and is still evolving [194], [195]. Applications for FTIR have seen a steady increase since its inception and they are many and diverse:

- analysis of human biofluids [196] like blood serum [197];
- biological [198], biochemical and biomedical analysis [191];
- identification of microorganisms [199];
- discovery of diagnostic and prognostic markers in cancer stem cells [200] [201];
- imaging of protein aggregation in living cells [202];
- drug resistance [203];
- online bioprocesses monitoring [204];
- preservation work [205];
- characterization of leucocytes [206];
- wine characteristics evaluation [207];
- flavonoids analysis [208] and green tea's EGCG extracts [209];
- (…).

2.6. Pre-processing methods and Multivariate Data Analysis

Spectra Pre-processing Methods

Spectral data pre-processing is a crucial step in any IR spectra acquisition and analysis, involving a well-defined and specific set of procedures which are performed on the raw data (the spectra), clearing the way for subsequent data mining tasks. Practical examples of the use of the advantages of pre-processed data in supervised and unsupervised environments, can be seen in chapter five. There are many reasons as to which data pre-processing is important regardless of the science field or object of study. These are summarized in Table 2.6.1.

Table 2.6.1. Highlights of pre-processing spectral data.

Robustness	A more robust and accurate analysis of the data and consequently of the classification and quantitative models.
Interpretation	It allows the user to be able to interpret and assign real world meaning to features, that would be otherwise non-interpretable, both by human beings and machines alike. The transformation of raw data into pre-processed data gives way for meaningful supervised and unsupervised analysis.
Outliers	Enabling the identification of outliers.
Reduction	Reduction of dimensionality in data and the selection of important features and removal of those that are not.

Atmospheric Correction

Water and carbon dioxide are known drawbacks for IR spectroscopy as these atmospheric gases can hinder the accuracy of the data and contribute negatively to the proper identification and quantification of the target analytes [210]. It is important to note that the fingerprint region is also affected by atmospheric components, mostly by H_2O, contributing to a downgrade in the ability to accurately quantify clinical parameters [211]. In order to avoid spectral interferences from atmospheric components, it is quite standard to have IR instrumentation in which both the interferometer and the sample chamber are closed and purged with dry air or N_2. However, this approach is not without its problems, as changes in purge performance or operator error can contribute to issues down the line.

Baseline Correction

Baseline correction (BC), also known as background correction, is a very important part of pre-processing. Various phenomena (e.g., fluorescence), can induce uneven amplitude shifts across different wavenumbers. These shifts need to be compensated before proceeding further with analysis. Any acquired spectra, measured by IR or other techniques, are bound to contain undesirable elements as noise and background in addition to the desired signal. As such, baseline correction methods are often used in infrared spectroscopy to adjust the spectral offset [176]. This is achieved by adjusting the spectral data to a minimum point in the data, effectively producing flat parts in the spectra (baselines) in-between well-defined peaks of absorption. There are a few baseline correction methods for FTIR spectra, such as **rubber-band**, **adaptive iteratively reweighted penalized least squares** (airPLS), **automated iterative moving average** (AIMA), **morphological weighted penalized least squares** (MPLS) and **iterative average** (IA). Regardless of being one of the newer methods and the one that seems to present the best results [212], the most traditionally used baseline correction method is still Rubber-Band. The way it works is as follows. The spectrum is divided in *n* ranges of equal size, where *n* is the number of baseline points. In the case of absorbance spectra, the minimum *y* value of each range is determined. All the minimum values are then connected by straight or spline (polynomial) lines, creating a baseline. From the x-axis, an imaginary rubber band is erected and stretched over the data curve. This rubber band is the baseline and points that do not lie on it are then promptly discarded.

Normalization

Normalization is used as a means to scale samples in order to get all the spectral data on approximately the same scale. A normalization is often applied in cases where the signal is a function of sample mass, for example, in most gas chromatography (GC) detectors or to compensate for intensity variations of the source power (e.g., IR synchrotron [213]), or due to different contents of the sample being analyzed. The normalization methods can be grouped in two. The first group refers to the simpler normalization methods that only require the information from the spectrum, whereas the

second group comprises of more complex normalization methods which require reference spectra. We refer to Table 2.6.2, regarding some normalization methods.

Table 2.6.2. Normalization methods used in the spectral data pre-processing.

	Name of method	Brief description
1st group normalization	**Peak Normalization**	Scales spectra in regards to a chosen peak or band [214], usually to Amide I band, at around 1650cm^{-1}.
	Standard Normal Variate (SNV)	SNV is a weighted normalization method whereas not all spectral data contributes equally. It is a row-oriented transformation which removes scatter effects from spectra by centering and scaling individual spectra [215]. Here, the average and standard deviation of all the data points for spectra is calculated. Every data point of the spectra is subtracted from the mean and divided by the standard deviation. It is sometimes used in combination with de-trending (DT) to reduce multicollinearity, baseline shift and curvature in spectroscopic data.
2nd group normalization	**Multiplicative Scatter Correction** (MSC)	MSC, also known as multiplicative signal correction was designed to deal with multiplicative scattering in reflectance spectroscopy. These unwanted scatter effects are then removed from the data matrix prior to data modelling. It is comprised of two steps. First, the estimation of the correction coefficients (additive and multiplicative) and lastly the correction of the recorded spectrum [216].
	Extended Multiplicative Scatter Correction (EMSC)	EMSC works in a similar way to MSC, but in the EMSC model a polynomial (generally a second order polynomial) is included with MSC, therefore being called Extended MSC. In addition, it allows for compensation of wavelength-dependent spectral effects, performing best when there is a prior knowledge of the unwanted spectra, which can then be subtracted [217].

Derivatives and the Savitzky-Golay filter

Derivatives are widely used in spectroscopic applications [218], to enhance spectral information, by resolving overlapping bands. It is important to note that derivatives cannot be performed with non-numeric data or when missing data is present in the data matrix. If so, steps should be taken into account to fill said missing values (usually be a mean of the neighboring values). In the case of FTIR spectrometers, it is possible to apply a Fourier-derivation. In this process, the spectrum is transformed into an interferogram and then multiplied by a weighting function. Finally, the spectrum is rebuilt to give the derivative. This gives a more sensitive result than normal derivation.

The first derivative is used to enhance the resolution of peaks, whereas the second derivative is helpful for more complex spectra where it can help to resolve overlapped bands. Derivatives can however result in reduction of the signal in the transformed data and noise amplification [219]. To minimize this, a *Savitzky-Golay* (SG) derivative or filter can be applied. SG was suggested in 1962 by Marcel J.E. Golay and Abraham Savitzky. SG can be applied up to fourth order derivatives. The SG

algorithm is based on performing a least squares linear regression fit of a polynomial around each point in the spectrum, effectively smoothing the data [220]. The derivative is then, for all effects, the derivative of the fitted polynomial at each point (with a chosen number of window points – *fitting*). The great advantage of this method is that the computation of the derivatives and the smoothing are both carried out in just one step [221]. The SG filter is most recommended for use in spectra containing a few sharp bands.

Multivariate Analysis

Multivariate analysis (MVA), a field of chemometrics, is used to extract information from very large and complex chemical and biological data sets and it involves the simultaneous analysis of a great number of variables [222]–[226].

Biological samples are, more often than not, comprised of large datasets. Not only that, given its intrinsic variability, along with the high dimension of FTIR spectra, transforms the data analysis into an impossibly time-consuming process. This is where the multivariate analysis comes into play as it allows for the reduction of dimensionality and complexity of the data and, consequently, extract any meaningful information. It has been determined that one important consideration when applying multivariate methods is the dataset size. The reason this is relevant is due to statistical significance. In the following sub-sections, it will be pointed out the main multivariate data analysis conducted in FTIR spectroscopy. To aid in the analysis of multivariate data there are supervised and non-supervised methods that look por patterns in data and perform data reduction, allowing for an easier interpretation. Such methods are briefly discussed in the following pages.

Principal Component Analysis

Principal Component Analysis (PCA) is one of the most used and powerful Exploratory Data Analysis (EDA) tools, using primarily a visual approach to find patterns in data [227]–[230]. Being the basic workhorse of many multivariate data analysis techniques, PCA has been long since extensively, used to study from the antibody activation of T-cells [231], [232] and spectral imaging [233] to the differentiation of embryonic stem cells [234], [235], among thousands of other studies. A solid understanding of this method is therefore considered a basic requirement for any data analysist.

PCA is a bilinear modeling method which aims to reduce the dimensionality of the data in order to properly describe the variation present in any given dataset. It provides an interpretable overview of the information in the data matrix. PCA is also known as a projection method, taking the information that is contained in the original variables and projecting it onto a smaller number of latent variables. These are designated as Principal Components (PC). Usually the first PCs represent a high percentage of variance in the database. However, it is important to note that, although the most variance is described indeed by the first two PCs, it is often the case (as we will see later) that the PC conjugations that best separate a given dataset in separate clusters may not be found in the two main PCs.

In the context of use of PCA in both The Unscrambler® X 10.5 and Orange software, each PC is described as a combination of *scores* and *loadings*. The principal component scores are described by the loading vector which is an explanation of the variance. In our spectroscopic context, the scores then represent values that correspond to a loading spectrum, which contains in itself a variety of peaks, positive and negative. This represents the spectral variation in our dataset. In a matrix representation, the PCA model, with any given number of PCs is represented by the following equation:

$$X = TP^T + E \text{ (**equation 1.1.**)}$$

In equation 1.1., we have represented the initial data matrix, where T pertains the scores matrix, P the loadings matrix and E the error matrix. Spectroscopic data is infamously known for having large data sets. PCA also enables the detection of patterns, clusters and subtle or gross outliers. Finally, it opens the way to quantify these differences, more effectively so when used in conjunction with other techniques such as Discriminant Analysis (DA).

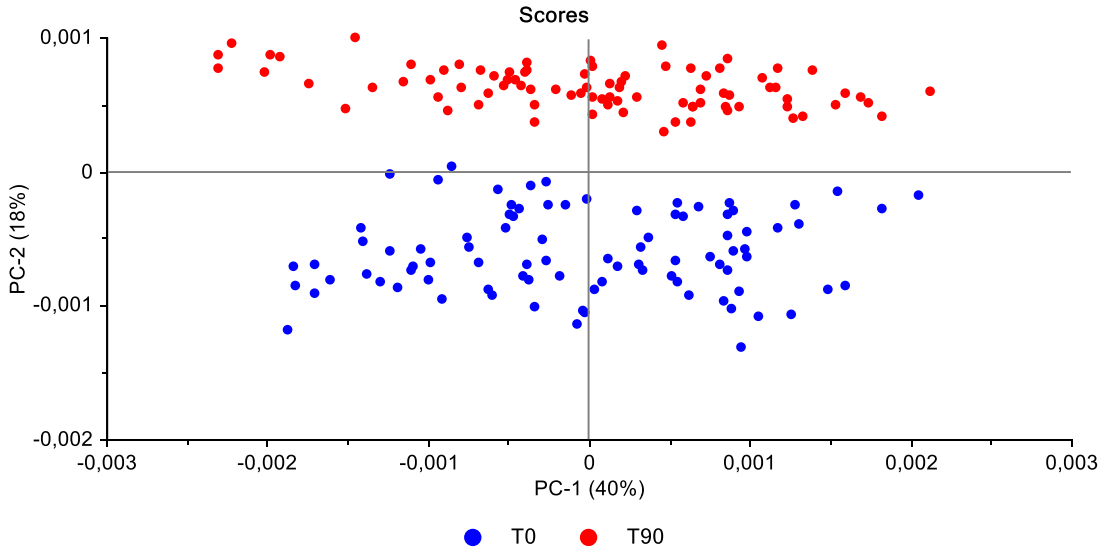


Figure 2.6.1. 2D visualization of principal component analysis scores plot for data obtained in FTIR plasma spectra, with a dilution factor of 10. Data is pre-processed with atmospheric correction and a second derivative, with a Savitzky-Golay filter, 2nd order polynomial and a 15-points window. In the figure, two separate clusters can be seen, T0 and T90. The first cluster representing the patients prior to the ingestion of the EGCG extract (T0) and the latter representing the same patients after ingestion of the green tea extract (T90).

Hierarchical Cluster

Hierarchical cluster analysis (HCA) is, similarly to PCA, an important tool for exploratory data analysis and classification of clusters [236]–[242]. It is best used in conjunction with other classification techniques, e.g., PCA and DA. The application of this EDA method is vast and well documented, having been used in the study of antibiotic sensitive and resistant microorganisms [243], discrimination of lactobacilli [244], analysis of foodborne pathogenic bacteria [245], pollen

characterization [246], human meningiomas [247], etc.

A *non-hierarchical clustering*, also called partitioning (e.g. K-means, K-medians) and *hierarchical clustering* (HC) methods can be applied, in which the results can be interpreted as *agglomerative* (bottom-up approach) or *divisive* (top-down approach). Agglomerative clustering fuses individuals (people, samples, etc.), into groups, whereas divisive clustering separates the individuals into finer groups. Agglomerative HCA is far more popular and it is the one used in this work. The non-hierarchical clustering methods allows to group things iteratively, based on their similarities in regards to specific characteristics (variables). It is important to note that this technique implies an *a-priori* knowledge of the data, as it is the user that decides from the start the desired number of clusters to separate the data. It is also important to mention that some methodologies are slower than others (K-medians), while others are more robust in regards to outliers (K-means).

HCA complete-linkage (also known as the farthest-neighbor method), uses the greatest distance between any two samples as its basis for the clustering. The Kendall's tau distance is more useful in identifying samples with a huge deviation in the data set, whereas the Spearman's rank correlation distance measures the correlation between two sequences of values. For hierarchical clustering methods, the dendrogram is the main graphical tool to get insight into the object of study. It is a tree-like display, in which the objects are clustered along the y-axis and the distance at which the cluster was formed is found along the x-axis. Distances along the y-axis have no meaning in a dendrogram. They are just equally spaced in order to make it easier to read. The way a dendrogram works is, basically, if one chooses any distance along the x-axis of the dendrogram and moves across the dendrogram, imagining a virtual line orthogonal to the x-axis, each line that is crossed represents a group or a cluster. Since the x-axis represents how close together observations were before being merged into a cluster, clusters that have branches very close together (in terms of distance), are usually not very reliable. However, if there is a big difference along the x-axis between two merged clusters, this is an indication that the clusters formed are most likely doing their job into displaying the real structure of the studied data.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is based on Bayes' theorem: $P(A/B) = \frac{P(B/A) - P(A)}{P(B)}$ [248], [249], which describes the probability of an event based on *a priori* knowledge of conditions that could be related to a specific event, e.g., the ingestion of an EGCG extract being able to change (or not), the serum metabolomics of an individual. Like so, one could, in theory, assess the probability that a person might have ingested the extract, based on the data provided from their clinical data or the analysis of the spectral data derived from the same.

LDA has been used in studies concerning medicine in general [250], bankruptcy prediction [251], [252], to decipher the hidden complexities of an ECG signal [253], osteoporosis identification [254], analysis of genetically structured populations [255], pattern and face recognition [256], [257],

hyperspectral data [258], etc.

It is the simplest of all classification methods. It provides a linear transformation of *n*-dimensional feature vectors, into an *m*-dimensional space, where this dimensional space *m* is inferior to our *n* samples ($m<n$). This way, samples that belong to the same class/group end up close together and at the same time separating T0 samples from T90. Their only difference being the position of their distribution centers (centroids). It works with the assumption that the probability distribution within the two groups are known and that the prior probabilities for the two are given and sum up to 100%.

It is also based on the assumption of a normal distribution and the assumption that the covariance matrices of our two groups are identical. Variance between and within groups is taken into consideration. In the event that there is no equal variance between groups and it is not possible to have a linear shape separating them, a linear curve might not be best suited and a quadratic discriminant analysis should be applied to provide a suitable classification model. This classification method falls under the *supervised* category, as the classes our samples fall in (and are to be classified), are known before the model is created, given the fact that our raw data already comes labeled from the start with the different timestamps (T0 and T90).

In the present work, the objective of LDA is be to determine the best-fit parameters for the classification of our samples and come up with a robust model, which can then be used to classify unknown samples. Usually with spectroscopic data, it is necessary to reduce the data dimensionality, e.g., through PCA, to overcome the constraint of LDA requiring more samples than variables in each class, giving way to PCA-LDA. Discriminant analysis may also be conducted by means of Partial Least Squares regression methods (PLS-DA) [259], with diverse applications as multivariate image analysis [260], metabolomics [261], etc.

PCA-LDA allows for a better visualization of cluster separation (and model accuracy), whereas in PLS-DA, the model accounts for the prediction of new and unknown samples in the data matrix and assesses its capability of identifying and separating into the previously formed clusters [262].

Regressions

Regression is a relatively old concept, first appearing in an article in 1885. Regression is a general term used for any and all methods that try to model and analyze a given number of variables with the sole purpose of finding a relationship between two separate groups of variables: dependent and independent variables. At this point, the fitted model can be used to describe said relationship or for prediction of new values. The objective of performing any regression is to discover how well *predictor variables* (X) can explain the variations in *response variables* (Y). In the spectroscopy field, our X variables are then represented by the spectra, whereas the Y variables have meaning as the individual constituents found within the spectra.

Noise can be a result of random variation in the response due to experimental error, measurement error, human error in samples preparation, etc., while irrelevant information is carried by predictors

which have little or nothing to do with the phenomenon we wish to study. For example, MIR absorbance spectra may carry information relative to the solvent and not only the compound(s) that we intend to study and determine their concentration. This is why it is important to use techniques in pre-processing that allow to overcome such issues like, e.g., normalization and second derivatives, which null the effect of concentration of the different compounds. As such, a good regression model is expected to model only relevant information, weighting it favorably, while downweighing irrelevant information. Lastly it should avoid overfitting, being able to distinguish between variations caused by the response (variation on the predictors) and variation caused by noise. To avoid noise in spectral data, spectral bands should be taken out of the equation altogether if deemed too noisy and irrelevant to the study.

As discussed before, exploratory data analysis is a first step in chemometric data processing and in some cases, simpler approaches like PCA might be all that is needed to characterize samples. However, due to its unsupervised nature, although providing a good picture and hopefully unbiased, of the data distribution, it lacks the possibility of providing predictions on new and unlabeled observations. This is a must for industries like the pharmaceutical where it is used in quality control and authentication of pharmaceutical products, by qualitative or quantitative methods, useful to quantify, for example, a specific compound (e.g., active ingredient) in a formulation [263].

There are many regression methods, such as Multiple Linear Regression (MLR), Principal Components Regression (PCR), Partial Least Squares Regression (PLSR), a variant of PLSR named L-PLSR [264]. Around 40 years ago MLR was the most sophisticated method available. It is a statistical procedure to predict the values of a response (dependent) variable from a collection of predictors (independent) variable values. While it has many applications, it has seen less use over the last years, due to some of their disadvantages (e.g., not being able to handle missing or incomplete data). It was also used at a time when the number of variables were much smaller than today. In 1970, PCR and PLS were first demonstrated in NIR applications. They are similar methods which use all the variables to form a smaller number of variables (named factors in PLS). There are not many choices in parameters regarding one or the other method, but PLS software developed faster and gained dominance in 1970, which is the reason why still today it is the method of choice when performing regressions, including in spectroscopy. Its advances in software allow for PLS to have traditionally better results than PCR. Not only that, it is also much faster, requiring a lesser number of factors for the same number of components needed in PCR, while maintaining or surpassing the covered variance. PLS does not come without is issues. As it uses all the variables in every factor, in any given calibration, some of the variables may be important in one factor but not in others, but all are included in every factor. What this translates to is that the influence of a variable on the prediction result, will ultimately depend on the size of the coefficient computed by PLS. Like so, if a coefficient is very small it will have a negligible effect on the result, but consequently it will also add a small noise component. As such, it would be ideal if it was possible to include variables in PLS that only made useful contributions to the precision of the predicted result, therefore diminishing the amount of noise in the

data. A relatively new method named Powered Partial Least Squares (PPLS), has been brought to light with promising results [265]. Ultimately PLS is still the most accessible and reliable method for fast results when dealing with spectroscopic data. It should be noted that for every single study, as long as enough samples are collected, their spectra are carefully measured and a suited calibration and validation is done, all methods described are still valid techniques, all with their advantages and shortcomings. Ultimately, calibration parameters for all methods should be compared (e.g., in a spreadsheet), such as **Standard Error of Calibration** (SEC), **Root Mean Square Error of Cross Validation** (RMSECV), **bias** (mean of the difference between the reference and predicted values) among others. Of the before mentioned regression methods we will only discuss in further detail PCR and PLSR, the regressions applied to this works' data.

Principal Components Regression (PCR)

PCR was first suggested as a solution to the multicollinearity problem in the late 70's by Greenberg [266], Fomby and Hill [267]. PCR was defined as "*a method of inspecting the sample data on design matrix for directions of variability and using this information to reduce the dimensionality of the estimation problem*". The use of principal components (PC) estimators as an estimating procedure in case of multicollinearity is attributed even earlier, in 1957, to Kendall [268] and later on to McCullum in 1970 [269]. Albeit the mouthful description from the original authors, PCR can be explained in a simpler fashion. It is a regression technique based on PCA, where the basic idea is to be able to calculate principal components (PC) and then use the necessary amount of them as predictors in a linear regression model – this is known as the typical least squares' procedure. It is a two step-procedure which first decomposes the X-matrix by PCA and then fits an MLR model using the PCs as predictors instead of the original X-variables.

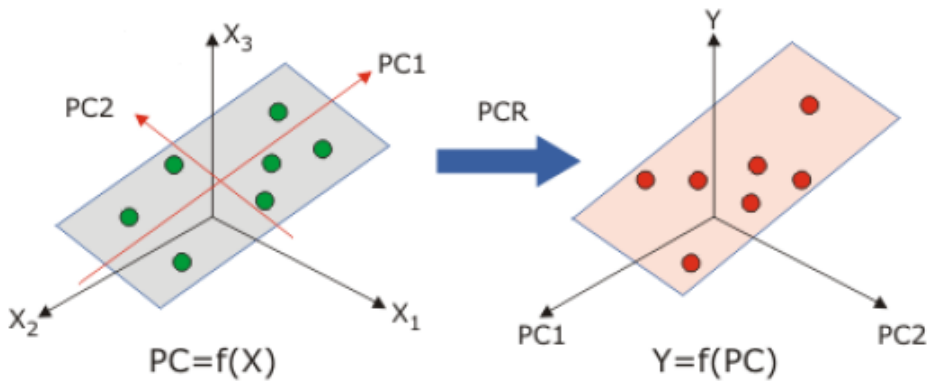


Figure 2.6.2. PCR procedure. In a first step, a PCA is done in order to find the orthogonal components. Then, an MLR model is fitted relating the PCs (X-variables: predictors) to the Y variables (response variables). ***(adapted from*** [270]***)***

PCR comes with a few notable advantages, to name a few: reduction of the dimensionality, partially mitigated overfitting and avoidance of the multicollinearity between predictors and consequently making the statistical inferences more reliable. PCR does not come without its pitfalls however, as it

is important to note that in PCR, it is assumed that the directions in which the X-variables (predictors) show the most variance (information) are the directions associated with the Y-variables (response variables). This is not 100% true every single time, but it gives the statistician or data scientist a very good approximation.

Partial Least Squares Regression (PLSR)

Partial Least Squares regression (PLSR), also known as Projection to Latent Structures (PLS), is an evolution of PCR in the sense that it models both X and Y-matrix simultaneously, in order to find the latent variables in X that will best predict the latent variables in Y. The components in PLSR are similar to those found in PCR, but here they are referred not as principal components but as *factors*.

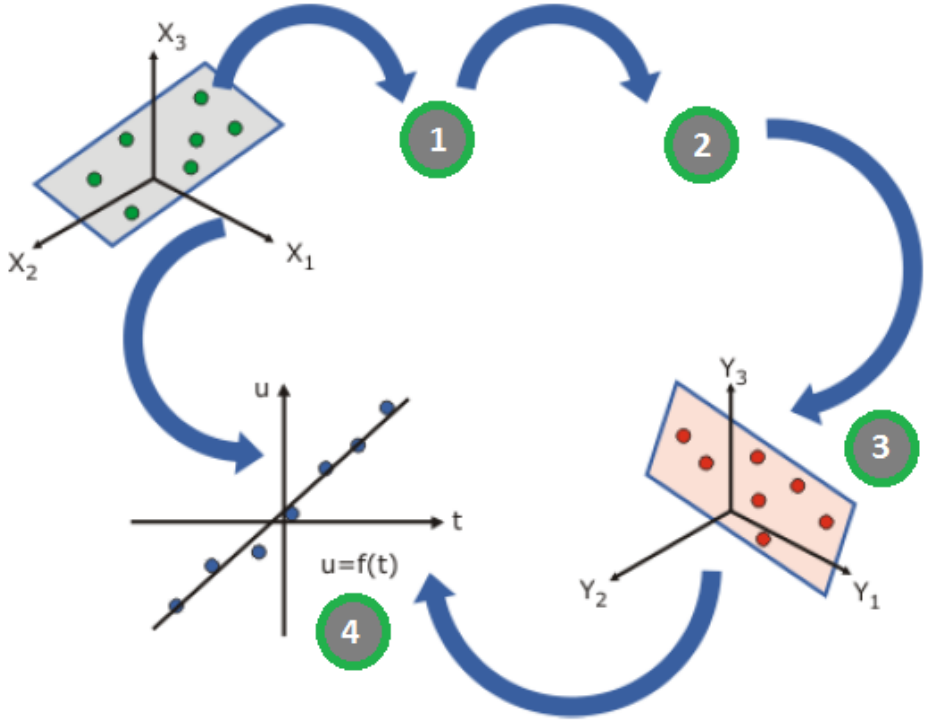


Figure 2.6.3. PLSR procedure. 1st step: the X scores (t) that are most correlated to Y are extracted; 2nd step: From (t), the Y-loadings (q) are generated; 3rd step: calculate Y-scores (u) from (q); 4th step: finally, both X-scores (t) and Y-scores (u) are plotted together in the same space, and their relationship is maximized. (adapted from [270]*)*

Not only does PLS regression provide, like other regression techniques, the ability to construct predictive models and test the significance of individual model parameters, through the simultaneous modelling of both X and Y variables, it allows its use in multivariate situations in which several response variables are collected and analysed at the same time, all the while being easily interpreted in a graphical visual display akin to PCA [271]. PLSR can consistently outperformed PCR in cases where there is strong collinearity in the data as well as when a great number of variables are present [272]. PLS regression, when dealing with spectral data, usually needs lesser components (factors), to reach a plus 90% variance, compared to PCR, leading to faster results [273]. This is observed in Figure 2.6.4.

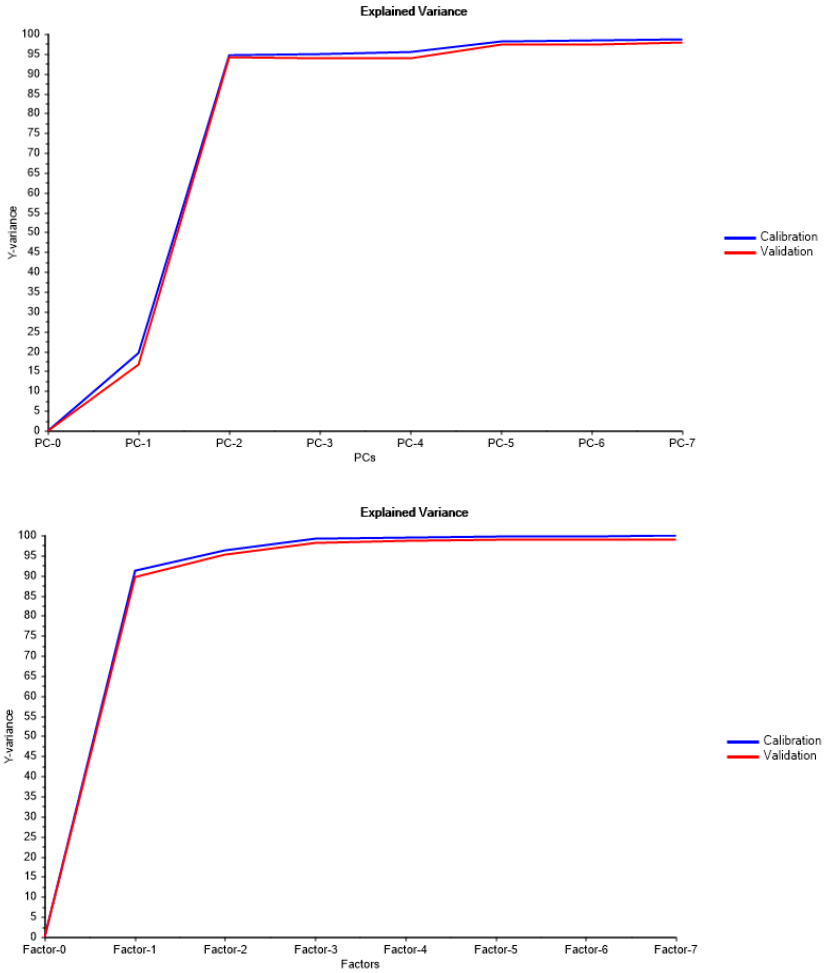


Figure 2.6.4. PCR (above) and PLS (below) representations for spectra of plasma pre-processed data with atmospheric correction and the second derivative, with a Savitzky-Golay filter (15 points window).

Chapter 3: Materials and Methods

This chapter briefly describes how the human clinical trials were performed for this work, regarding the acquisition of samples, how they were processed in laboratory, equipment used, as well as the experimental methodologies for the acquisition of FTIR spectra of human plasma and serum and subsequent spectra pre/processing and multivariate data analysis.

3.1. Equipment and solutions

The following main equipment and materials were used throughout this work.

Equipment

- FTIR spectrometer (Vertex 70, Bruker), with an HTS-XT module (Bruker);
- Centrifuge (Mikro with 1195/L rotor, Hettich);

- 96-wells Si micro-plate (Bruker);
- Vacuum pump (Vacuubrand, ME 2);

Solutions

- Sodium chloride (NaCl) from Sigma-Aldrich, at 0.9% (w/v);

3.2. Biological Assay

For this work, the effect of the consumption of EGCG along 90 days was evaluated in human volunteers (n=30). All participants provided a signed written informed consent before enrolment in the study, approved by the ESTeSL Ethics Commission. The blood samples were taken from 20 women donors and 10 male donors. All of them had extensive bloodwork analysis done (hemogram, leukogram, etc.) [274], [275].

Participants took, orally, 225mg of EGCG present in commercial capsules each day for 90 days. Two blood samples were taken just before the start of the assay (T0) and at the end of the 90 days (T90), respectively. In one tube, used to collect blood, Ethylenediaminetetraacetic acid (K3-EDTA) was added as an anti-coagulant agent. After the blood coagulated in the dry tube (i.e., without K3-EDTA), and the tube with K3-EDTA were centrifuged at 3500 rpm for 10m and the supernatant was extracted. The consequent plasma and serum samples were then stored and kept frozen at -20ºC.

3.3. Characterization of the human volunteers and its blood clinical analysis

A *t*-student statistical analysis (based on Microsoft's Excel), of 35 clinical variables from the 30 human volunteers over T0 and T90 was conducted.

3.4. FTIR Spectroscopic Analysis

MIR Spectral Acquisition

Triplicates of 25 μL of plasma and serum, diluted at 1/10 in water, were transferred to a 96-wells Si plate and then dehydrated for about 2.5 h, in a desiccator under vacuum. Spectral data was collected using a FTIR spectrometer (Vertex 70, Bruker) equipped with an HTS-XT (Bruker) accessory. Each spectrum represented 64 coadded scans, with a $2cm^{-1}$ resolution, and was collected in transmission mode, between 400 and 4000 cm^{-1}. The first well of the 96-wells plate did not contain a sample and the corresponding spectra was acquired and used as background, according to the HTS-XT manufacturer.

Spectral Data Analysis

A review of all the spectral pre-processing and processing techniques, as well as the varied multivariate data analysis methodologies used for both chapter 4 and 5 can be found in Figure 3.4.1 and Figure 3.4.2.

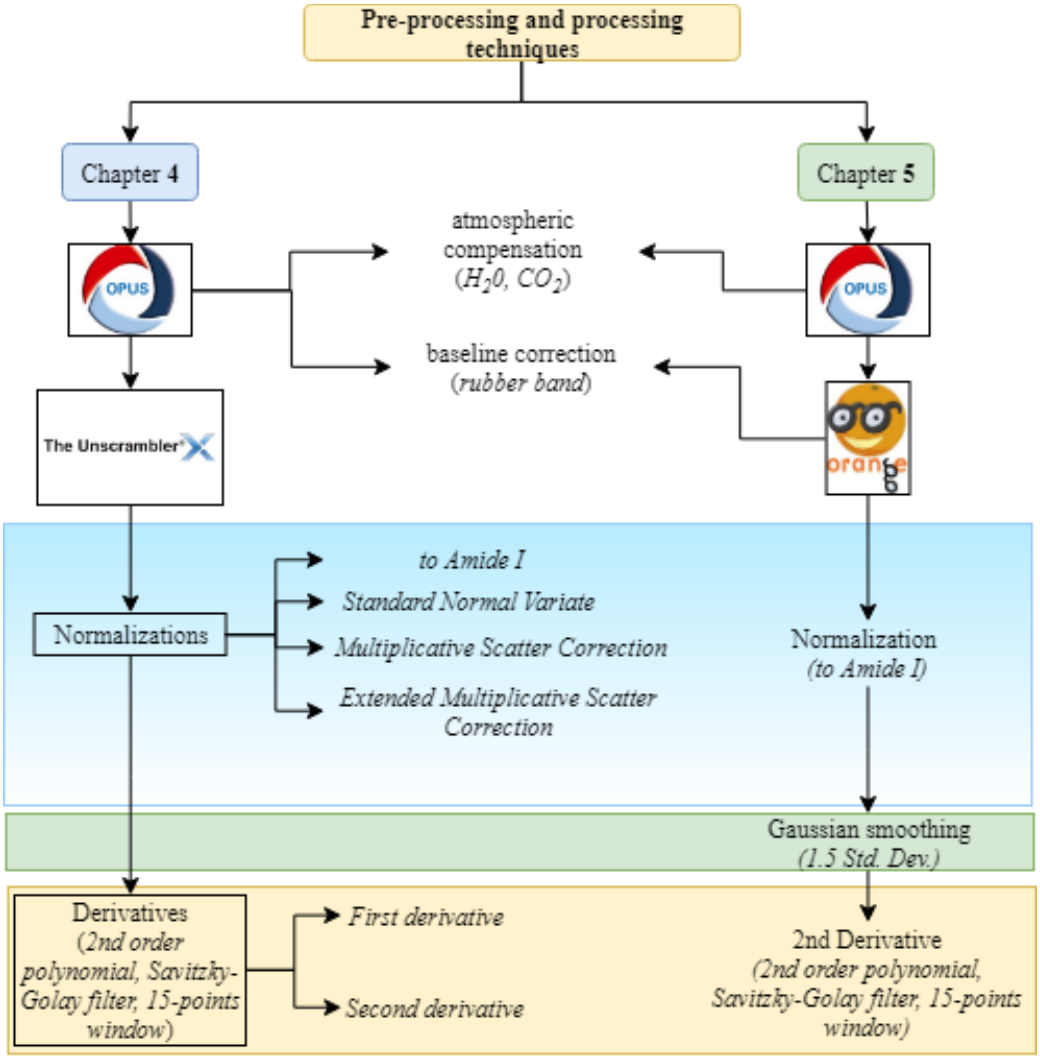


Figure 3.4.1. Pre-processing techniques and respective software used in chapters 4 and 5.

On Figure 3.4.2 is represented the multivariate data analysis performed on chapters 4 and 5. To note that the presented diagram is not meant to reflect the actual order of the techniques used, since for example, in Chapter 5, where the Orange open source software is used, the order of applied techniques (exception being to data pre-processing), is mostly irrelevant as everything is calculated at the same time. As such, the way the diagram is presented is done so in a way to facilitate comparison between the methods used in chapter 4 and 5, where similar techniques are applied with e.g. PCA, HCA being highlighted in similar colors, while unique and exclusive methods used in either chapter 4 or 5 being displayed in a different color pallet.

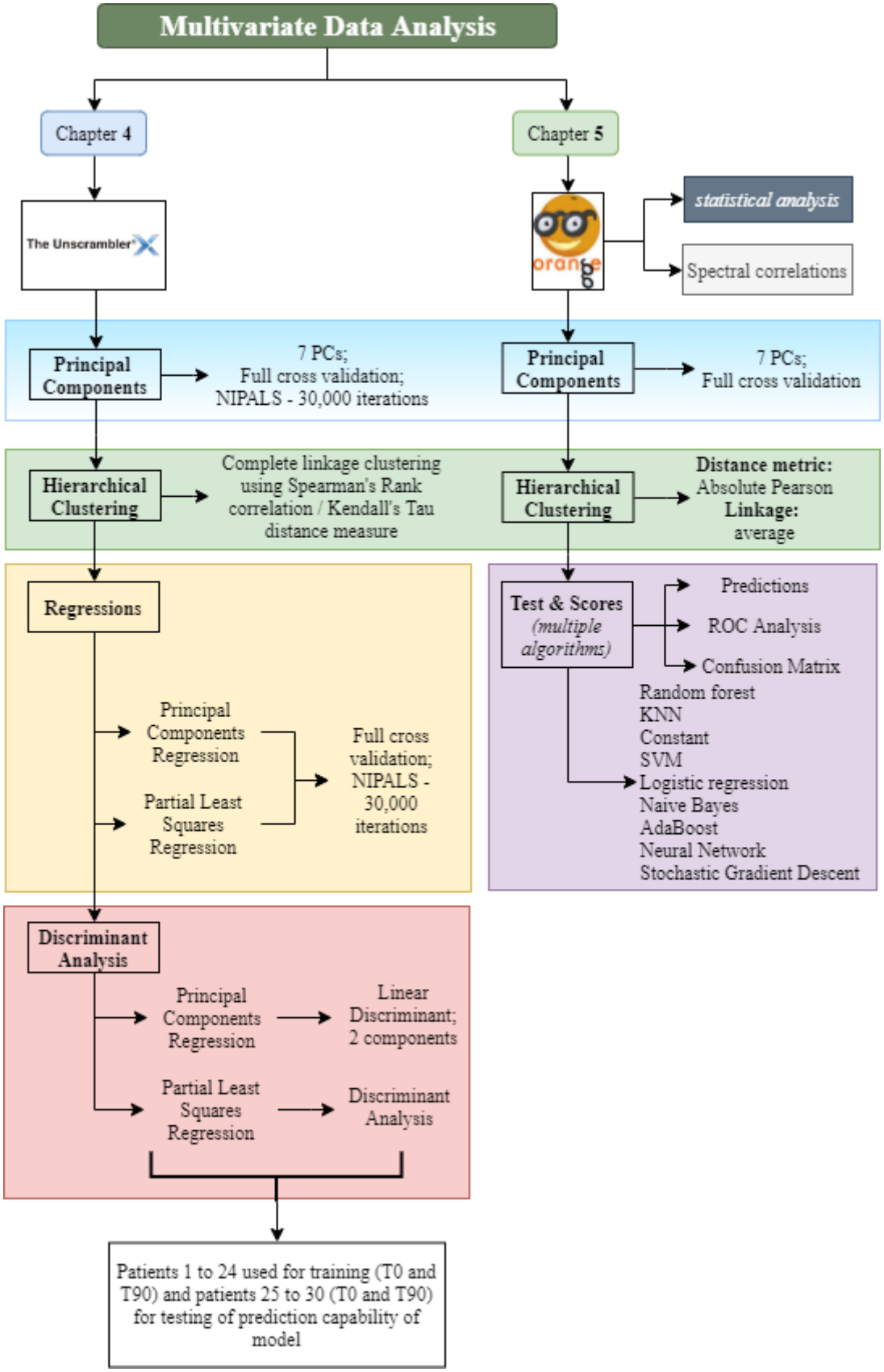


Figure 3.4.2. Multivariate Data Analysis and corresponding software used in chapters 4 and 5.

Chapter 4: Results and Discussion

Tea has become a staple in many western countries as of late, in part due to the ever-increasing number of reports concerning its health benefits, such as prevention of potentially pathological changes (e.g., high cholesterol), aiding in weight loss or due its anti-inflammatory capabilities due to neutralization of free radicals. Previous studies in human clinical trials have established that 400 mg to 800 mg of EGCG intake are considered safe doses [276]. These doses result in peak serum concentrations in the range of 100 to 400 ng/ml [277]. To assert on the safety and capabilities of the most abundant catechins in green tea, a clinical study comprehending 30 voluntary participants was conducted. It included the daily consumption for 90 days of a bolus dose of green tea extract (GTE), containing 225 mg of EGCG. Blood samples were taken from all participants just before the start of the EGCG intake, designated as time 0 (T0) and immediately after the last bolus intake, designated as time 90 (T90). These samples were clinically analyzed as described in subchapter 3.3. The main goal of the following subchapters was to evaluate the impact of EGCG intake after 90 days, i.e. by comparing analysis at T0 and T90, on the molecular fingerprint of the participants plasma and serum, analyzed by FTIR spectroscopy.

Chapter 4 briefly analyzes the conventional blood clinical tests, focusing subsequently on the main pre-processing and processing techniques to obtain information from the FTIR spectra acquired from the plasma and serum. The following pre-processing techniques were evaluated: atmospheric compensation, baseline correction, normalization (to Amide I, Standard Normal Variate), derivatives (first and second). Regarding unsupervised and supervised classification methods were applied: PCA, PCR, PLSR, DA.

4.1. Characterization of the human volunteers and its blood clinical analysis

Thirty individuals have taken part in the study, of which 20 were women and the remaining 10 were men. The clinical blood analysis conducted over the 30 participants at T0 and T90 are summarized in Table 4.1.1. It is indicated the average and standard deviation for each of the 35 types of blood analysis conducted at T0 and T90, respectively. From the 35 clinical blood analysis, 7 were observed to be statistically different (at a 5% confidence level), between T0 and T90 (Table 4.1.1). These 7 are represented as box plots in Figure 4.1.1.

*Table 4.1.1. Clinical blood analysis conducted at the 30 participants at T0 and T90, respectively and its corresponding average and standard deviation values. The p-value of the t-test comparing T90 and T0 are represented, being highlighted in **bold p-values lower than 5%**.*

Variable		T_0		T_{90}		***p*-value** (<0.05)
		Mean	**Standard Deviation**	**Mean**	**Standard Deviation**	**2 tail (Paired)**
1	Erythrocytes	4,5730	0,3130	4,6930	0,4107	**0,017667942**
2	Hemoglobin	13,7967	1,1822	14,2900	1,3484	**0,001831928**
3	Hematocrit	40,3100	2,9420	41,9300	3,9424	**0,000807382**
4	Mean cell volume	88,1000	3,4175	89,3667	3,3268	**2,54675E-05**
5	Mean Corpuscular Hemoglobin	30,1000	1,8071	30,4333	1,4065	0,105706265
6	Mean corpuscular hemoglobin concentration	34,2333	0,9714	34,3667	1,0334	0,5362031
7	Red Cell Distribution Width (RDW)	12,9467	0,7234	12,8233	0,6590	0,08971563
8	Leukocytes	7,0817	1,5221	7,4833	1,4297	0,09789263
9	Neutrophils	4,2823	1,1979	4,5853	1,1858	0,14927143
10	Neutrophils %	59,9367	7,6595	61,0833	7,9396	0,46378631
11	Eosinophils	0,1853	0,1402	0,1760	0,1005	0,66405144
12	Eosinophils %	2,6767	1,9942	2,3800	1,3823	0,26015037
13	Basophils	0,0850	0,3108	0,0300	0,0064	0,34034736
14	Basophils %	1,3667	5,2947	0,4000	0,0000	0,32558199
15	Lymphocytes	2,2017	0,6870	2,3643	0,7043	0,20804905
16	Lymphocytes %	32,1733	6,7844	31,7233	7,6376	0,74703447
17	Monocytes	0,3340	0,0800	0,3277	0,0946	0,66952221
18	Monocytes %	4,8133	1,1866	4,4133	1,0608	0,08145051
19	Platelets	243,1000	53,8807	251,3667	54,8688	0,17247869
20	Reticulocytes	1,5867	0,3989	1,4467	0,4455	0,09245809
21	Absolute value	72,3000	18,4749	68,8333	20,1222	0,32338795
22	Reticulocyte hemoglobin content	31,9667	2,0424	33,2333	1,6955	**7,62521E-05**
23	Fetal hemoglobin (HbF) level	0,3967	0,1402	0,4500	0,1570	**0,000182746**
24	Triglycerides (mg/dL)	107,7667	45,0721	119,1667	41,4388	0,08625708
25	Triglycerides (mmol/L)	1,2280	0,5137	1,3590	0,4725	0,08465204
26	Total cholesterol (mg/dL)	176,7333	41,8692	175,0000	39,0384	0,63719699
27	Total cholesterol (mmol/L)	4,7503	0,9813	4,7053	0,8049	0,63741258
28	HDL (mg/dL)	61,6000	13,8180	63,5333	14,8504	0,13038322
29	HDL (mmol/L)	1,5950	0,3594	1,6457	0,3853	0,12768864
30	LDL (mg/dL)	100,3000	28,5852	94,3000	22,4532	0,10282363
31	LDL (mmol/L)	2,5963	0,7390	2,4559	0,5866	0,17004306
32	LDL/HDL	1,6794	0,4935	1,5607	0,5367	**0,042125114**
33	Aspartate aminotransferase	24,2333	8,8967	25,4000	9,2684	0,545971048
34	Alanine aminotransferase	25,3667	8,3851	30,7333	22,0109	0,172069204
35	Gamma-glutamyltransferase	19,1333	18,8217	21,8667	25,4338	0,167181629

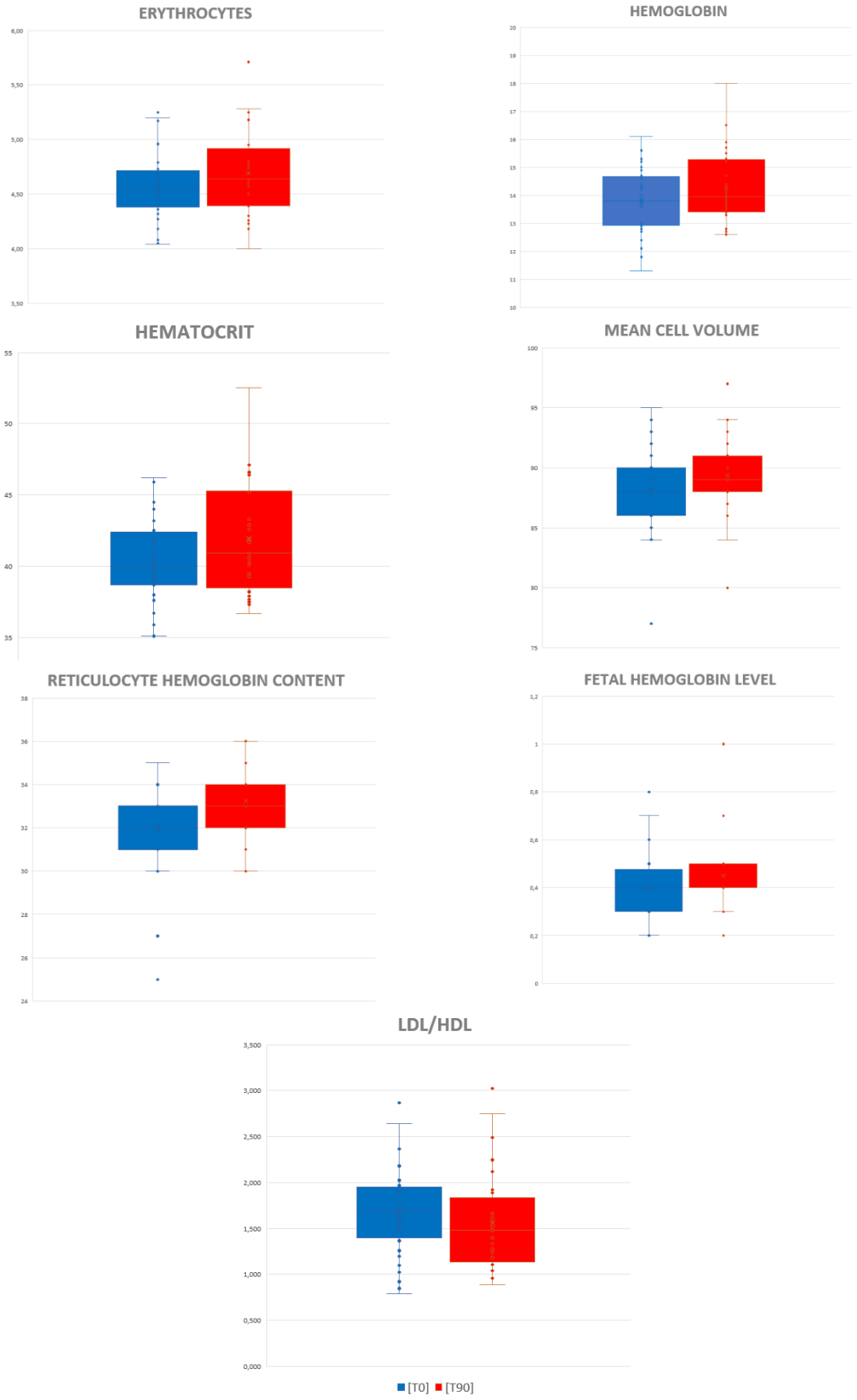


Figure 4.1.1. Boxplots of 7 blood analysis of the 30 participants at T0 and T90. The last three blood analysis presented moderate outliers.

4.2. Optimization of pre-processing methods and identification of outliers

Plasma samples: optimization of spectral pre-processing

To minimize physical interferences while enhancing chemical and biological data on FTIR spectra, diverse spectra pre-processing techniques were evaluated by PCA. Figure 4.2.1 to Figure 4.2.3 represent triplicates of plasma spectra diluted with water at 1/10, of 30 participants at T0 and T90, with diverse pre-processing and the corresponding PCAs. Based on the observations of these figures, the best non-derivative pre-processing, that which enabled the best separation between T0 and T90 samples on the PCA scores plot, was a combination of atmospheric and baseline correction. The second derivative pre-processing was the best derivative to contribute to the separation of T0 and T90. Of these, the second derivative pre-processing was the best at contributing to the separation of T0 and T90.

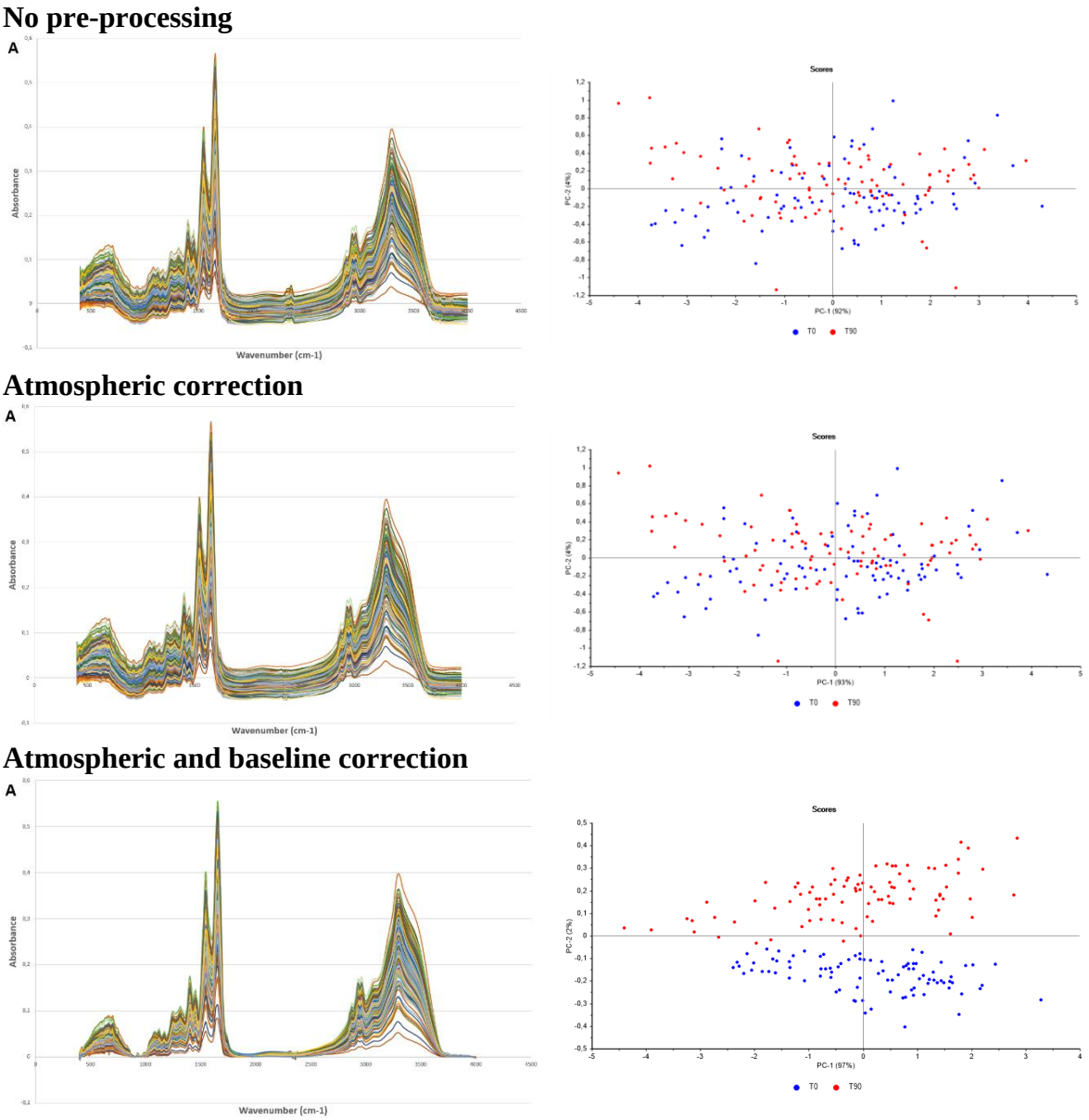
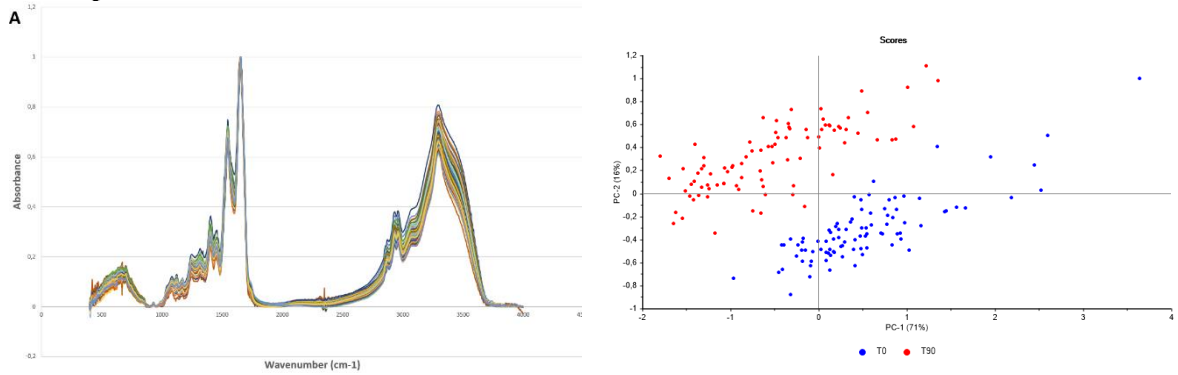
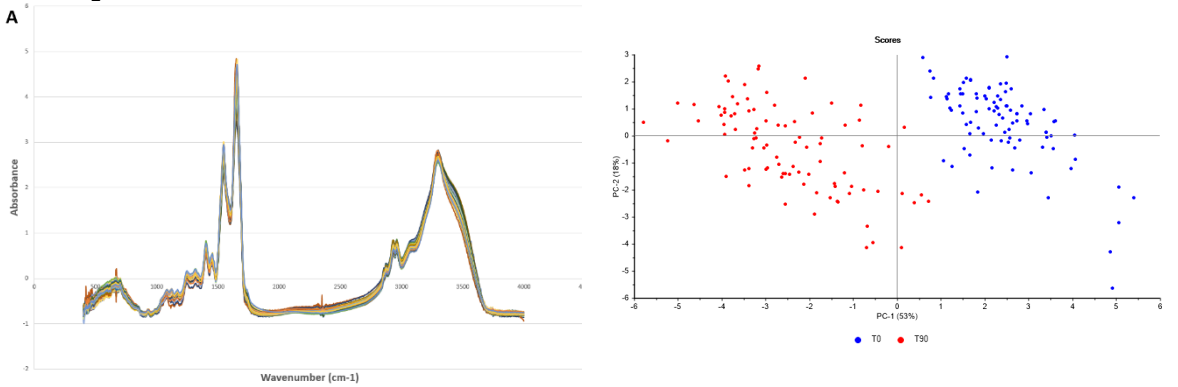


Figure 4.2.1. Left: FTIR spectra of plasma diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red.

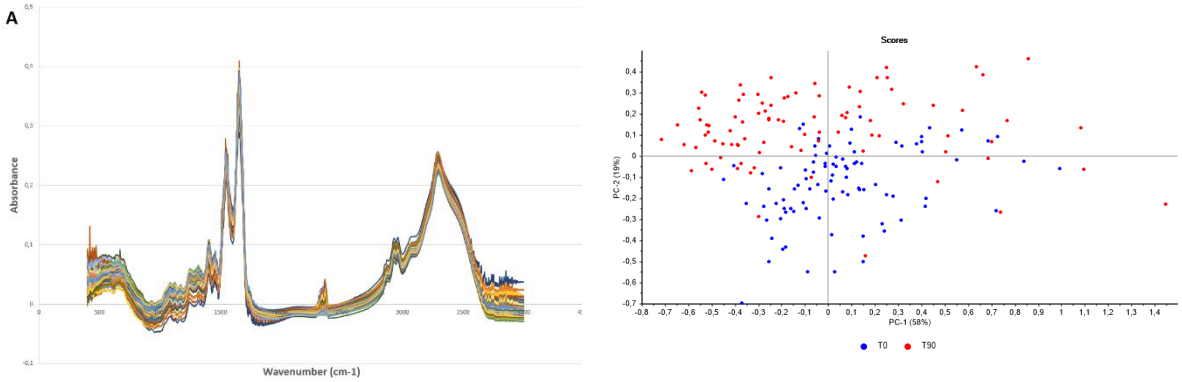
Atmospheric and baseline correction and Normalization to Amide I



Atmospheric and baseline correction and SNV



MSC



Extended MSC

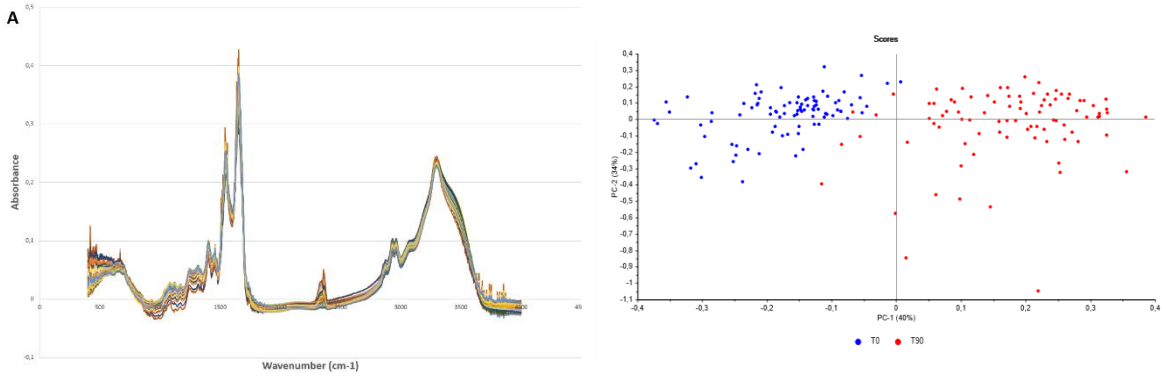
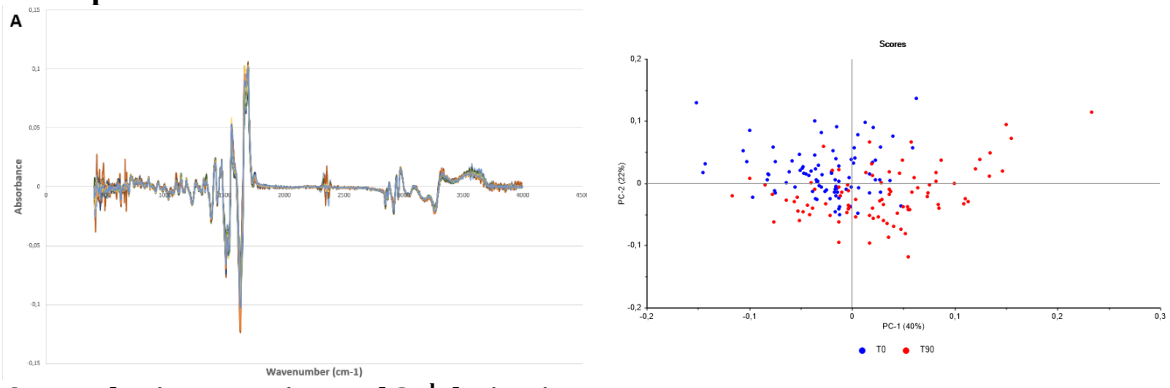


Figure 4.2.2. Left: FTIR spectra of plasma diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red.

Atmospheric and baseline correction and SNV and 1st first derivative



Atmospheric correction and 2nd derivative

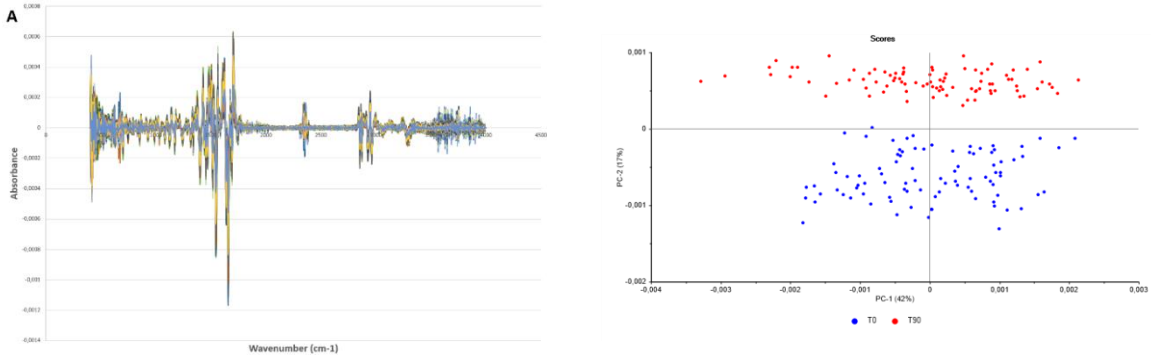


Figure 4.2.3. Left: FTIR spectra of plasma diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red. The derivative was based on a Savitzky-Golay filer with a 2nd order polynomial and a 15-points window.

Plasma samples: identification of outliers

For the best pre-processing techniques – atmospheric and baseline correction, atmospheric correction followed by the 2nd derivative, the spectral outliers were identified taking into consideration the explained variance, F-residuals in function of Hotelling's' T2 and Hotelling's ellipse both at 1% significance. The identified outliers are discriminated in Table 4.2.1. Afterwards and with the outliers already eliminated from the spectral data, a new analysis was conducted. A more detailed graphical visualization of the non-processed plasma spectral data, along with the two best pre-processing techniques, before removal of outliers, can be seen in Figure 4.2.4 through Figure 4.2.6.

Table 4.2.1. Samples identified as outliers in triplicates of FTIR spectra of plasma diluted to 1/10 from 30 participants acquired at T0 and T90 and pre-processed by atmospheric and baseline correction or by atmospheric correction followed by 2nd derivative.

Pre-Processing	Patient	Time (days)	Replicate number deleted	Designation
Atmospheric and baseline correction	**1**	90	1	P1-T90-1
	3	90	3	P3-T90-3
	10	90	3	P10-T90-3
	12	0	2	P12-T0-2
Atmospheric correction and second derivative	**1**	90	1	P1-T90-1
	3	90	3	P3-T90-3
	7	90	1	P7-T90-1
	10	90	3	P10-T90-3
	24	0	2	P24-T0-2

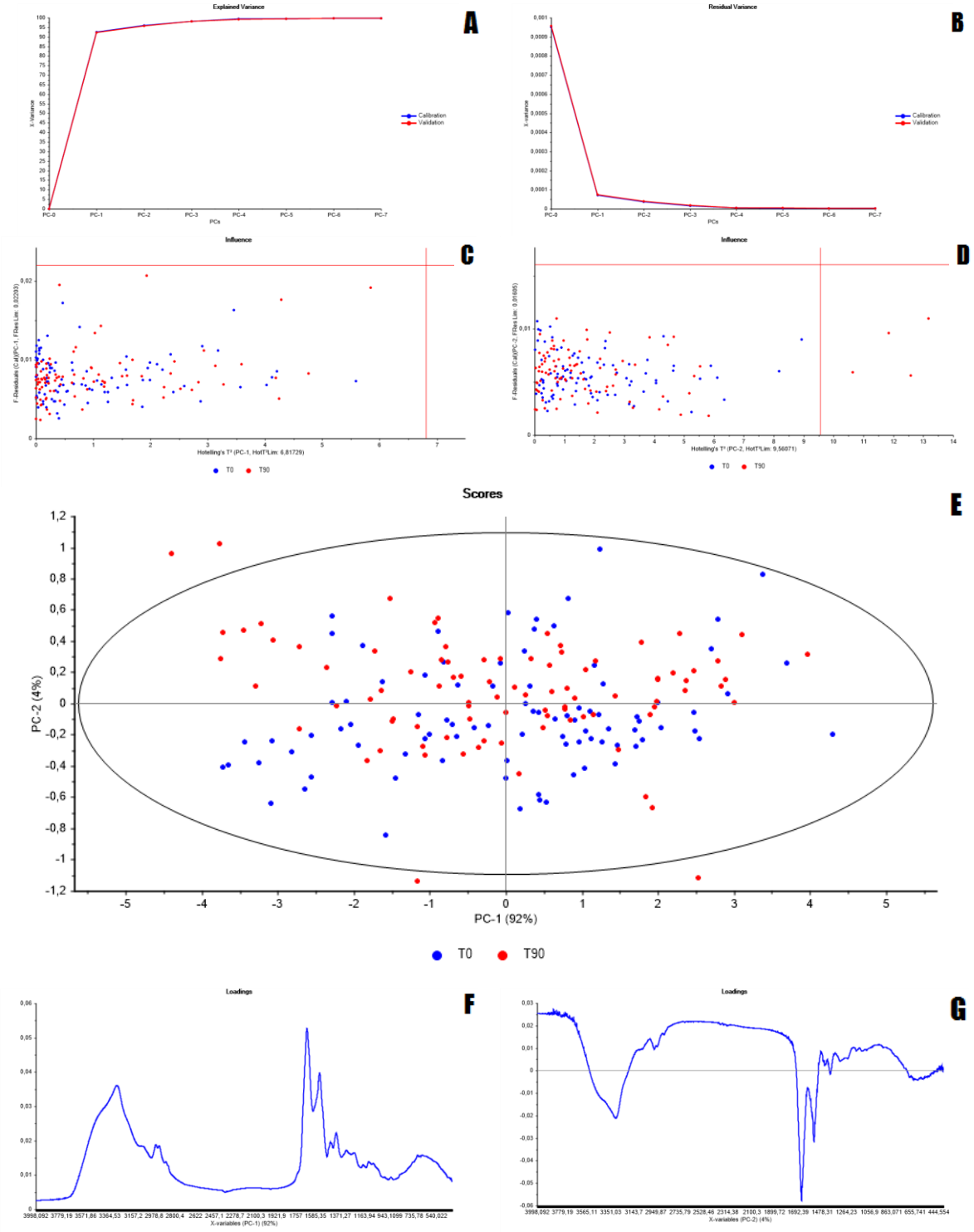


Figure 4.2.4. Explained variance (A) and residual variance (B) for the raw, unprocessed plasma spectra; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E) for PC1 versus PC2; loadings for PC1 (F) with 92% variance and PC2 (G) with 4% variance.

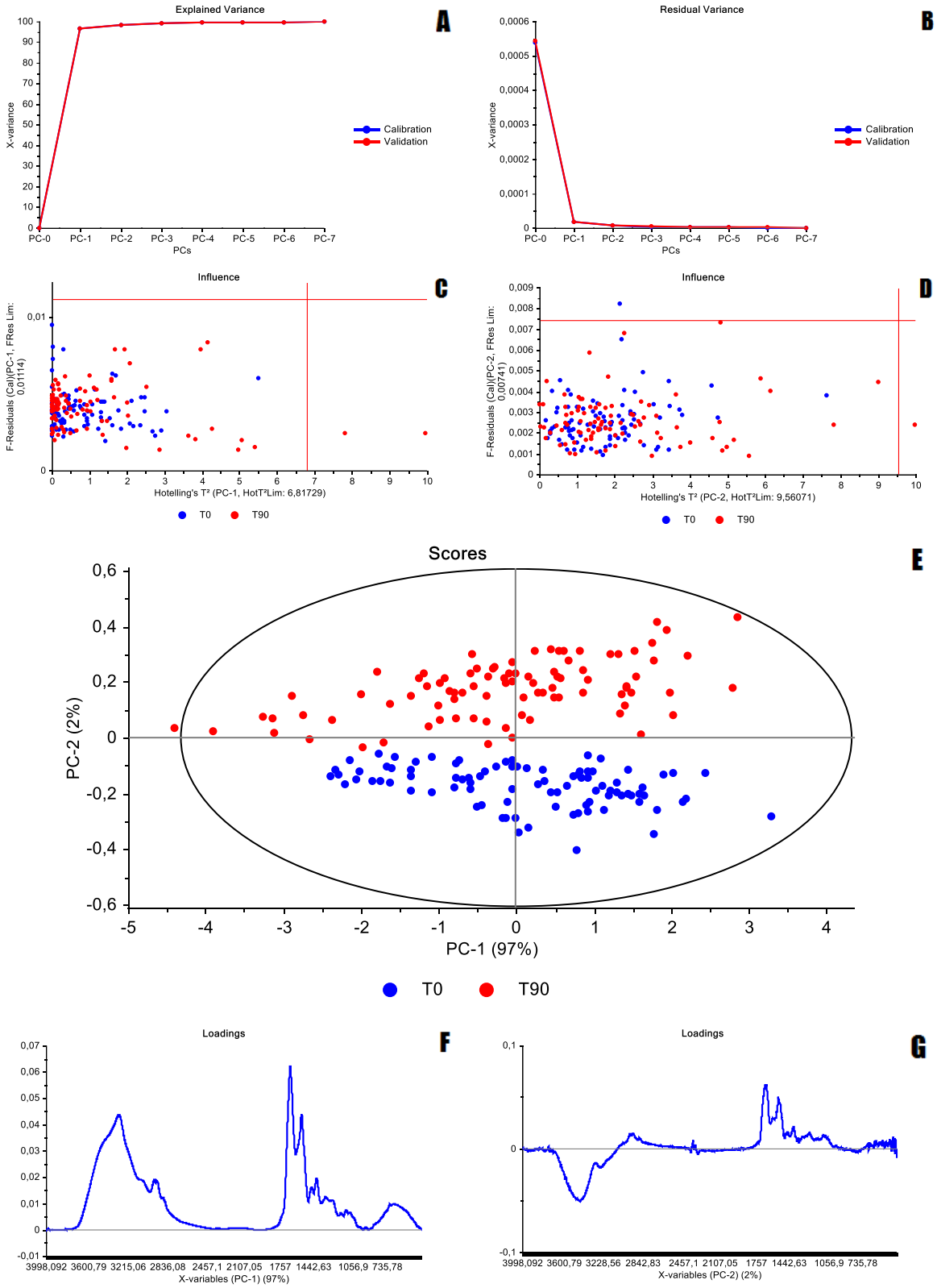


Figure 4.2.5. Explained variance (A) and residual variance (B) for the pre-processed plasma spectra with atmospheric and baseline correction; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E) for PC1 versus PC2; loadings for PC1 (F) with 96% variance and PC2 (G) with 2% variance.

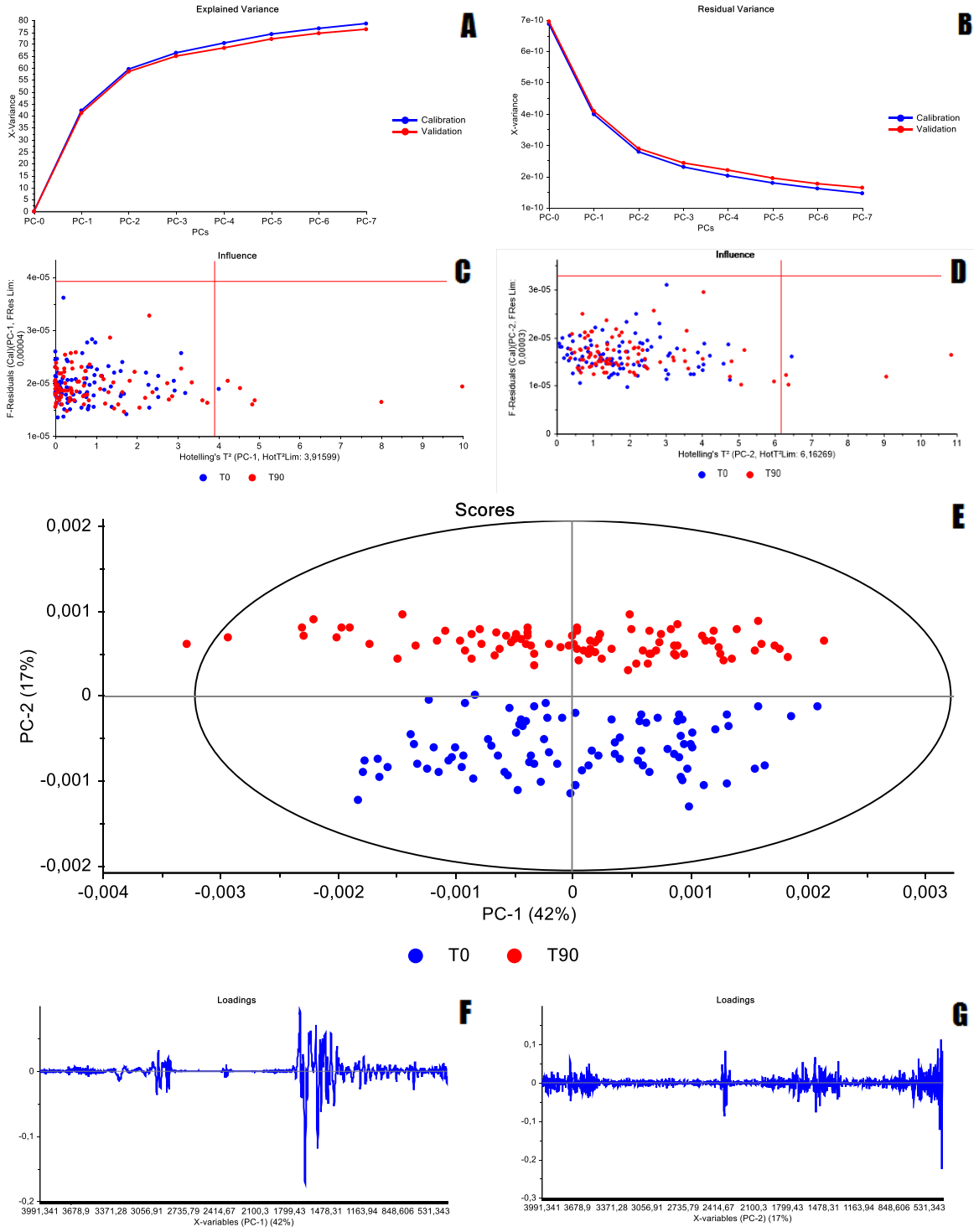


Figure 4.2.6. Explained variance (A) and residual variance (B) for the pre-processed plasma spectra with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15 smoothing points; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E); loadings for PC1 (F) with 40% variance and PC2 (G) with 18% variance.

Serum samples: optimization of spectral pre-processing

Figure 4.2.7 to Figure 4.2.9 represent triplicates of serum spectra diluted with water at 1/10, of 30 participants at T0 and T90, with diverse pre-processing and the corresponding PCAs. Based on the observations of these figures and similarly to what had happened to plasma before, here too the best non-derivative pre-processing, that which enabled the best separation between T0 and T90 samples on the PCA scores plot was a combination of atmospheric and baseline correction. The second derivative pre-processing was the best derivative to contribute to the separation of T0 and T90.

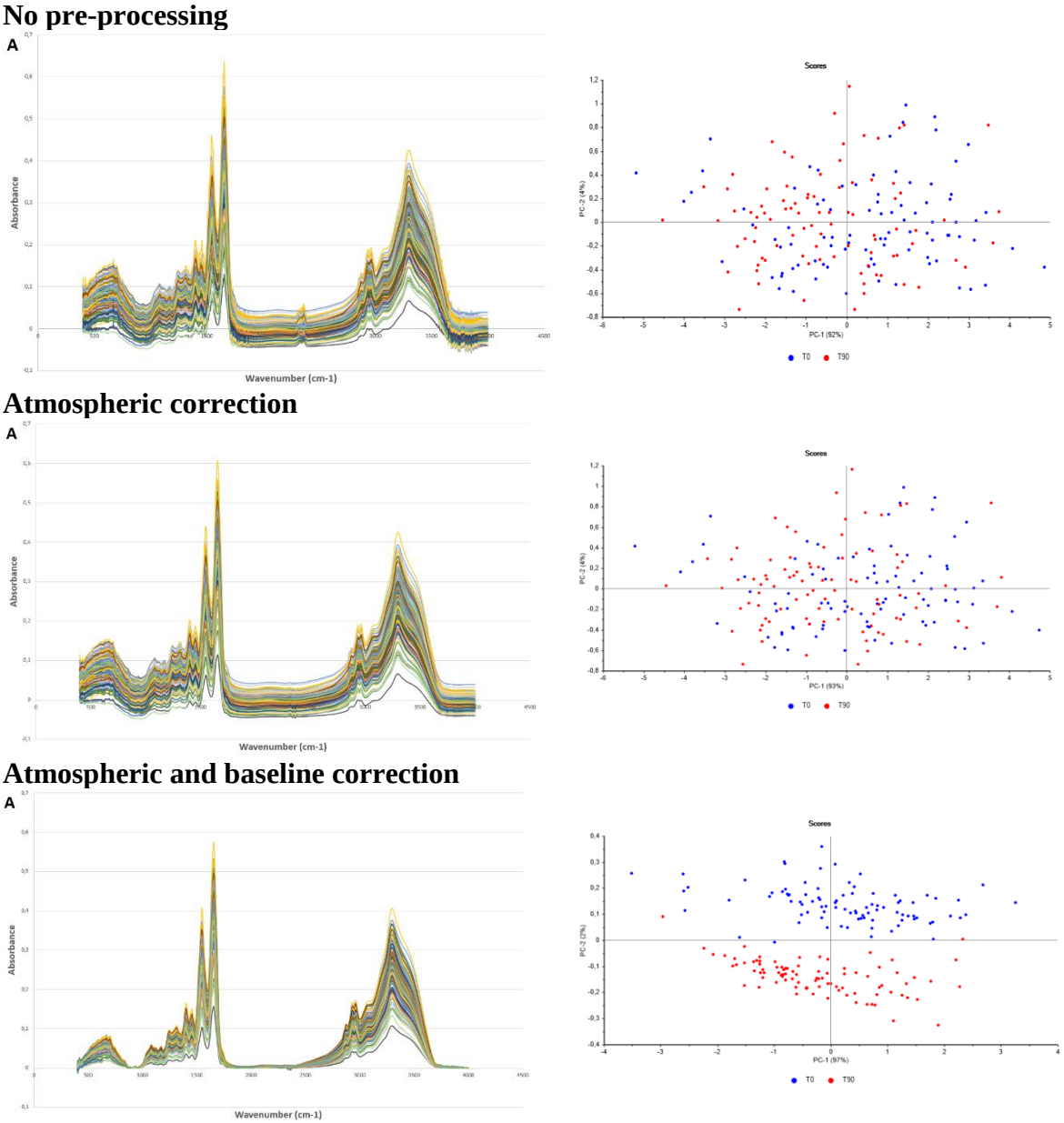
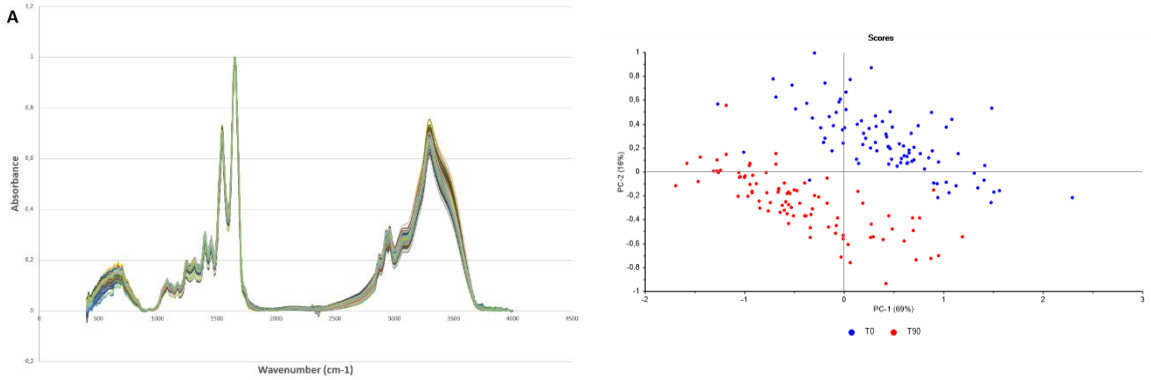
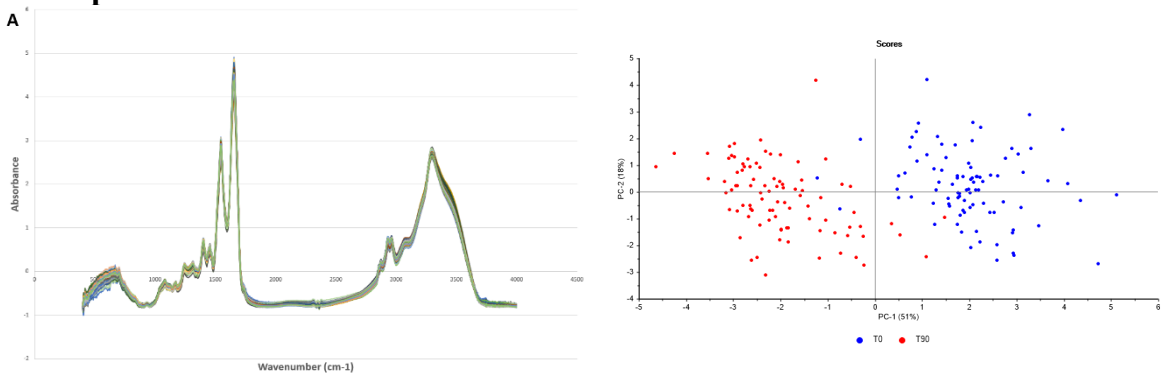


Figure 4.2.7. Left: FTIR spectra of serum diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red.

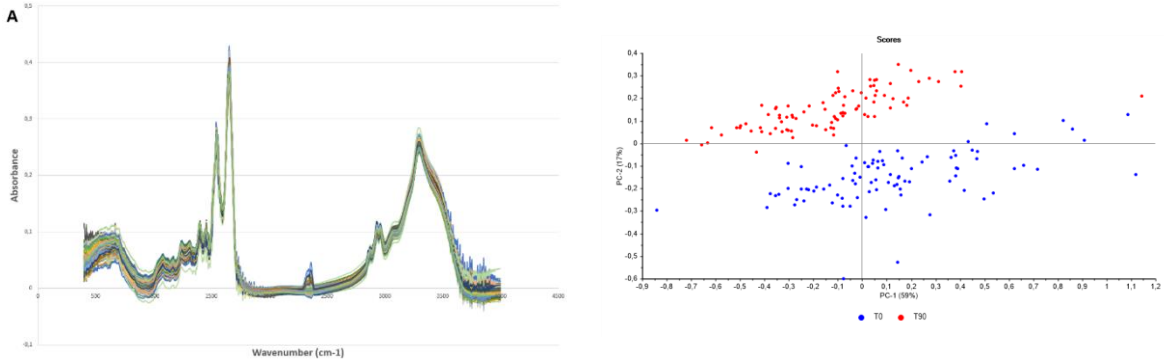
Atmospheric and baseline correction and Normalization to Amide I



Atmospheric and baseline Correction and SNV



MSC



Extended MSC

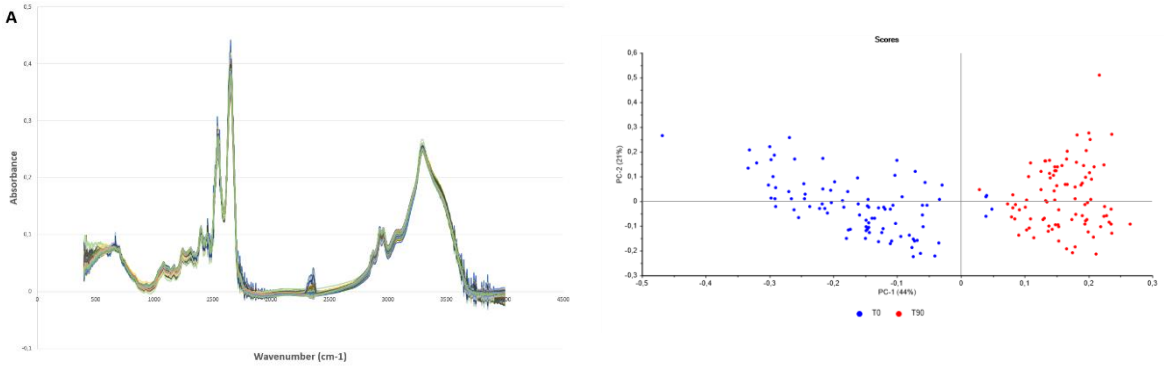
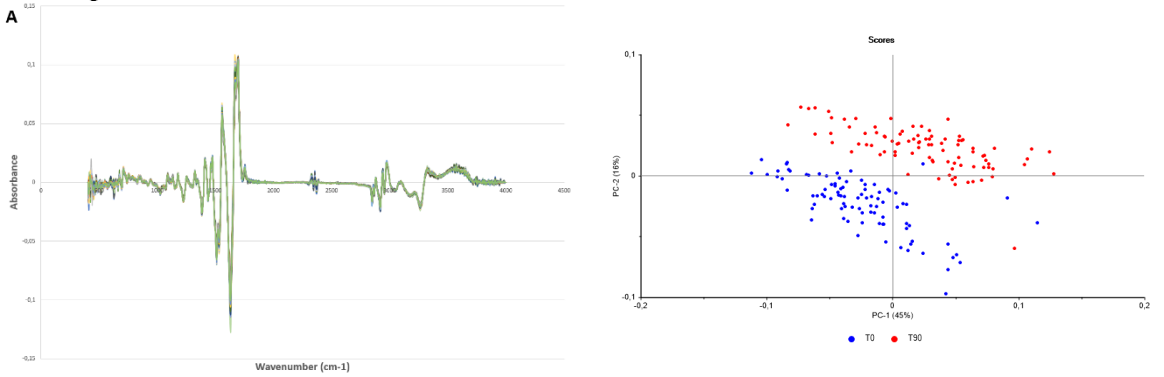


Figure 4.2.8. Left: FTIR spectra of serum diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red.

Atmospheric and baseline correction and SNV and 1st Derivative



Atmospheric Correction and 2nd derivative

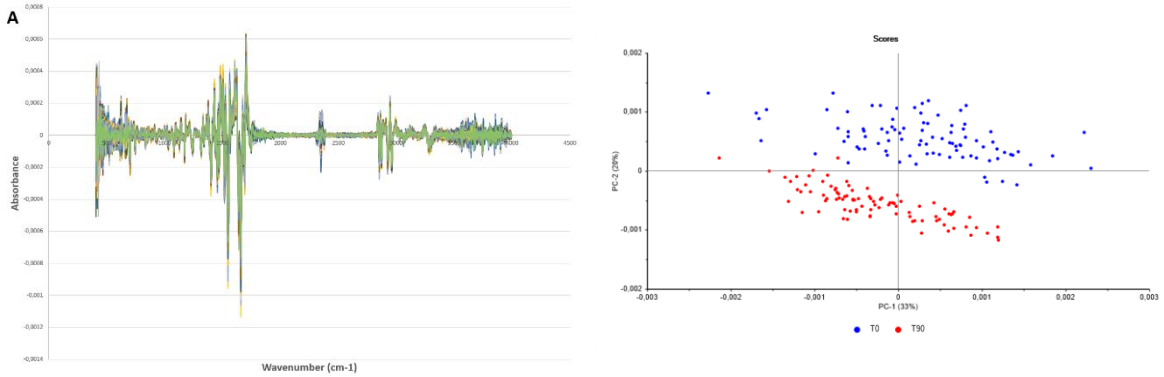


Figure 4.2.9. Left: FTIR spectra of serum diluted to 1/10, with diverse pre-processing techniques. Right: corresponding 2D PCA, with representation of T0 in blue and T90 in red. The derivative was based on a Savitzky-Golay filer with a 2nd order polynomial and a 15-points window.

Serum samples: identification of outliers

Regarding the identification of outliers for serum samples, similarly to what was done for plasma, they are discriminated in Table 4.2.2 and the graphical visualization of the non-processed serum spectral data, along with the two best pre-processing techniques, following outlier removal, can be seen in Figure 4.2.10 through Figure 4.2.12.

Table 4.2.2. Samples identified as outliers in triplicates of FTIR spectra of serum diluted to 1/10 from 30 participants acquired at T0 (in blue) and T90 (in red) and pre-processed by atmospheric and baseline correction or by atmospheric correction followed by 2nd derivative.

Pre-processing	Patient	Time (days)	Replicate number deleted	Designation
Atmospheric and baseline correction	**24**	0	1	P24-T0-1
	29	0	3	P29-T0-3
Atmospheric correction and second derivative	**6**	90	3	P6-T90-3
	11	90	3	P11-T90-3
	24	0	1	P24-T0-1
	27	0	3	P27-T0-3
	28	0	2	P28-T0-2
	29	0	3	P29-T0-3

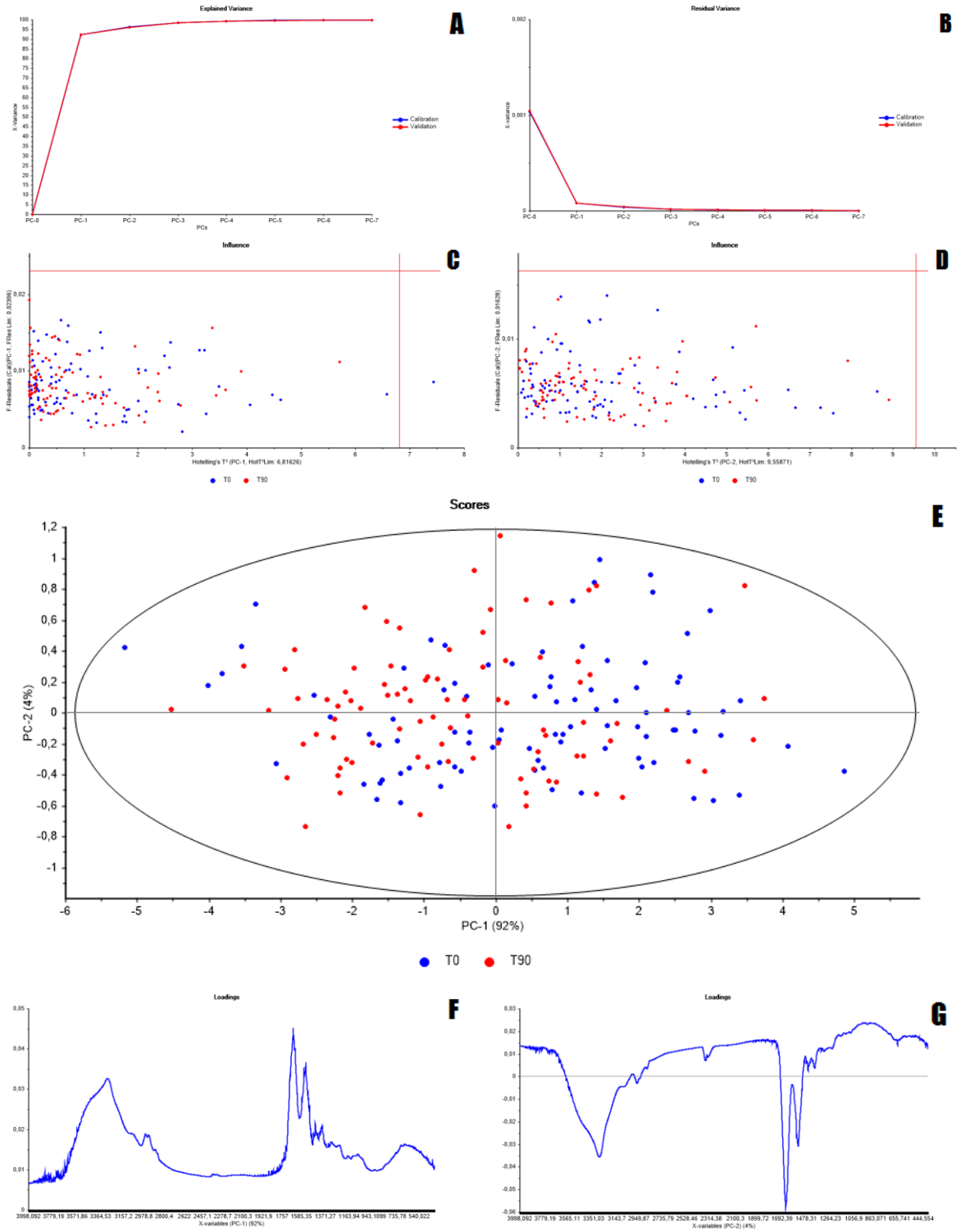


Figure 4.2.10. Explained variance (A) and residual variance (B) for the raw, unprocessed serum spectra; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E) for PC1 versus PC2; loadings for PC1 (F) with 92% variance and PC2 (G) with 4% variance.

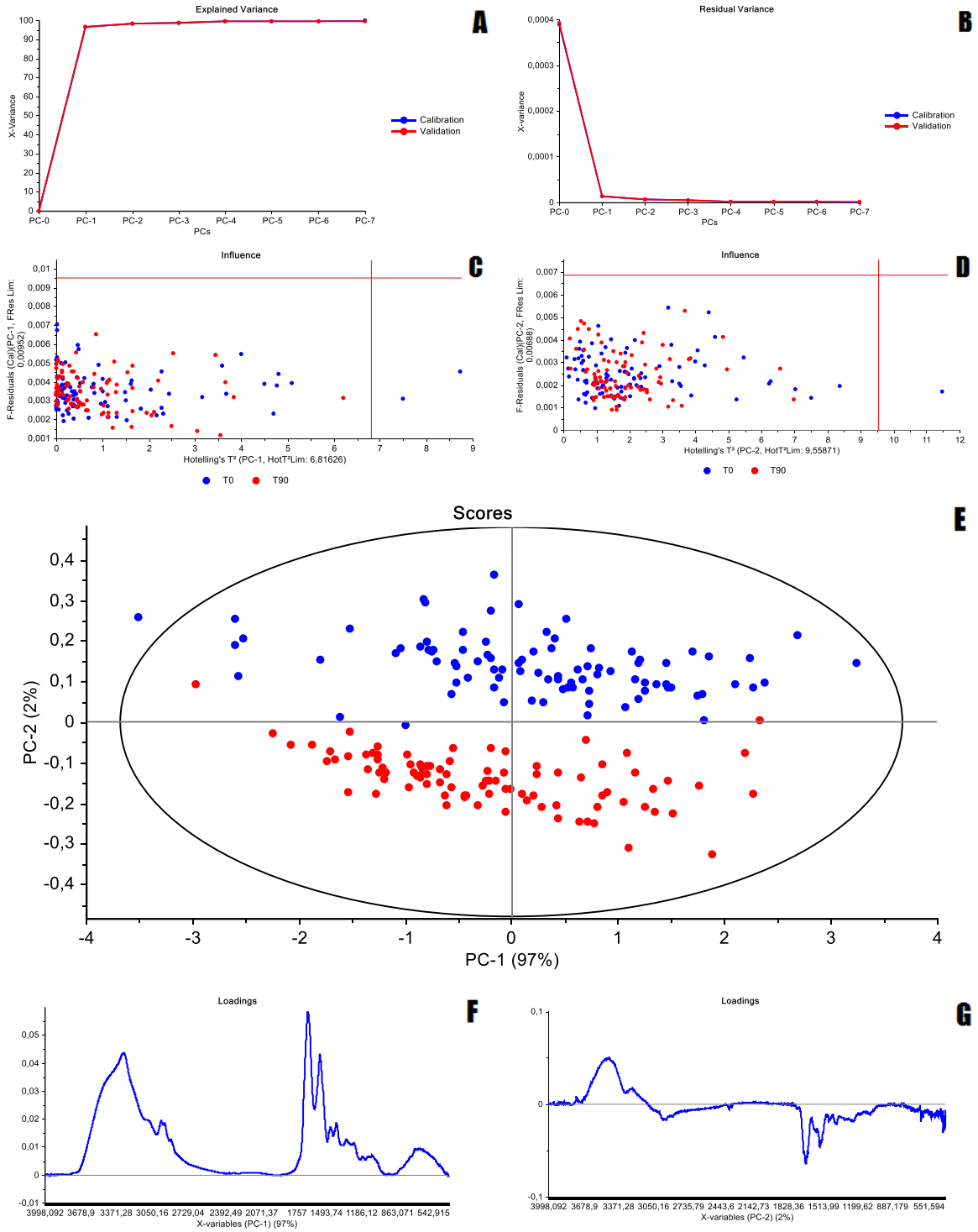


Figure 4.2.11. Explained variance (A) and residual variance (B) for the pre-processed serum spectra with atmospheric and baseline correction; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E); loadings for PC1 with 96% variance (F) and PC2 (G) with 2% variance.

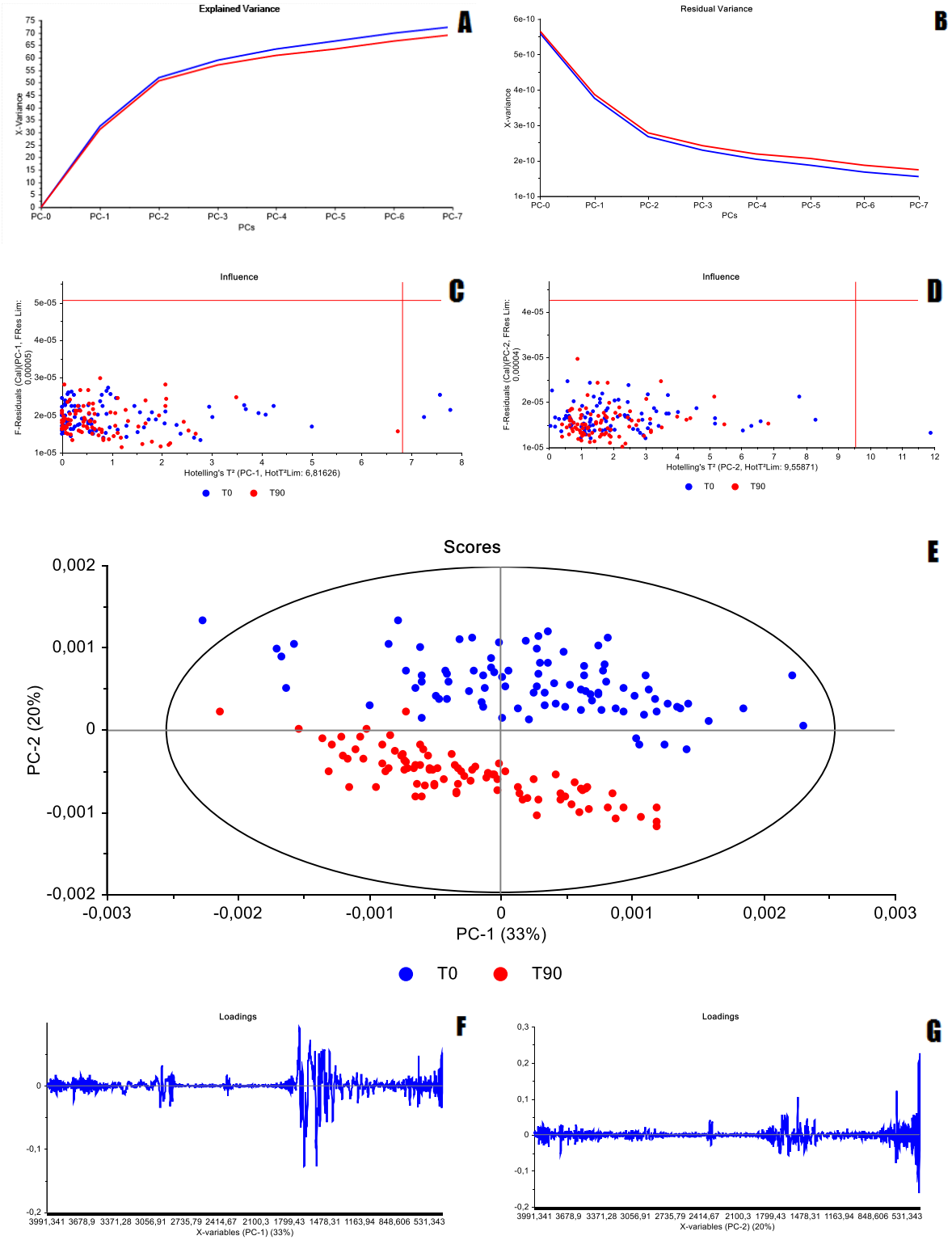


Figure 4.2.12. Explained variance (A) and residual variance (B) for the pre-processed serum spectra with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15 smoothing points; (C) influence diagram for PC1 and PC2 at 1% significance (D); scores diagram with Hotelling's T^2 ellipse at 1% significance (E); loadings for PC1 (F) with 30% variance and PC2 (G) with 21% variance.

4.3. Hierarchical Cluster Analysis

Hierarchical Cluster Analysis (HCA) was applied for non-processed spectra and pre-processed spectra by atmospheric and baseline correction or by atmospheric correction followed by second derivative. HCA was conducted either on the replicates of all samples (without elimination of outliers), or on reduced data, where the average spectra of the replicates were considered. The HCA Kendall's Tau distance or the Spearman's rank correlation were evaluated. Regarding the plasma samples, both T0 and T90 show a perfect separation in the pre-processing data of atmospheric correction and second derivative, for both triplicate and reduced data, as can be seen in the figure below.

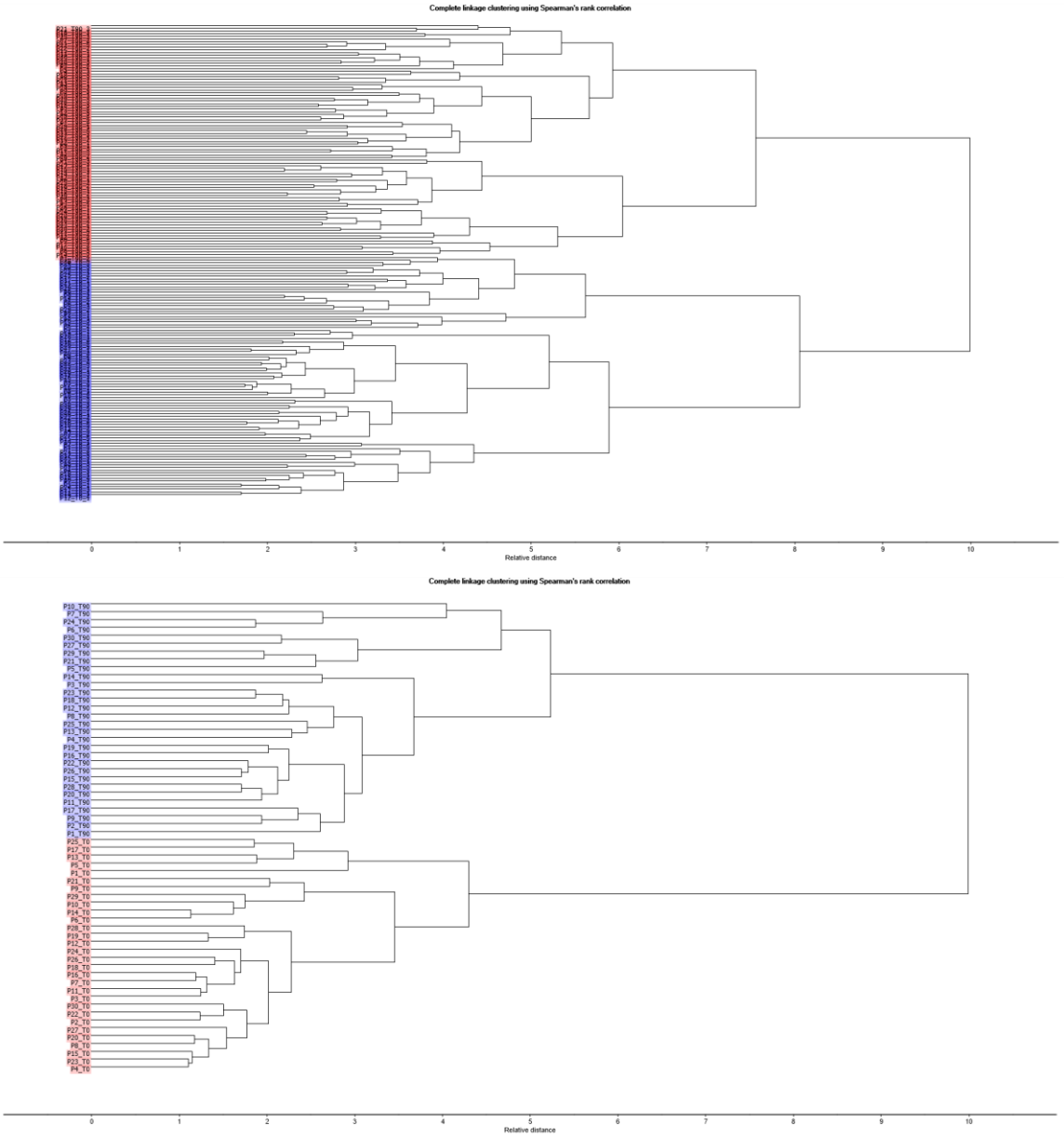


Figure 4.3.1. Hierarchical Cluster Analysis for the plasma spectra, T0 and T90, pre-processed with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15-points window. HCA complete linkage method and Spearman's rank correlation distance measure was used. Above: all three replicates are shown for each patient (T0 in blue and T90 in red). Bottom: replicates are reduced to a single value per patient (T0 in blue and T90 in red).

As can be seen for the case of serum, in Figure 4.3.2, in the HC where replicates are used, patient 10, at the beginning of the study (coded as **P10_T0)** is classified correctly in the T0 cluster. However, it displays an elongated branch of its own (marked in light blue and enveloped by a green box), showing it to be slightly different from the other samples at T0. The influence of this sample becomes evident in the bottom image of Figure 4.3.2, with the reduced data, where it distorts data in such way that it shows as a cluster of its own. This issue could have been avoided by doing outlier removal first and then either averaging the replicates or using median instead, but it was not done so, in order to get a better understanding of the data's behavior by means of HCA.

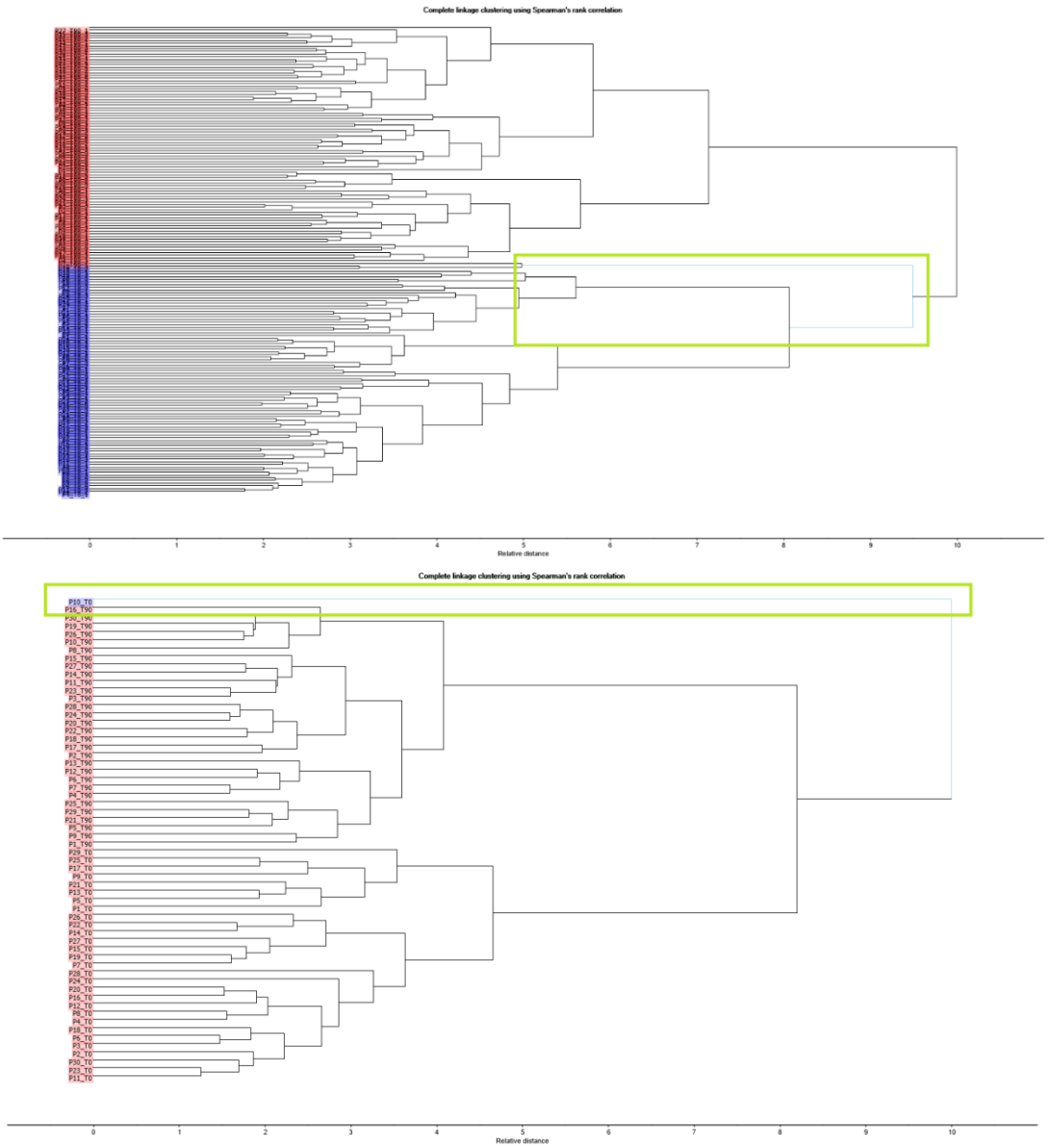


Figure 4.3.2. Hierarchical Cluster Analysis for the serum spectra, T0 and T90, pre-processed with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15-points window. HCA complete linkage method and Spearman's rank correlation distance measure was used. Above: all three replicates are shown for each patient (T0 in blue and T90 in red). Bottom: replicates are reduced to a single value per patient (T0 in blue and T90 in red).

To verify the weight that this P10_T0 sample has, the sample was taken out and a new HC was performed and can be seen in Figure 4.3.3, where T0 and T90 display a perfect and clear separation.

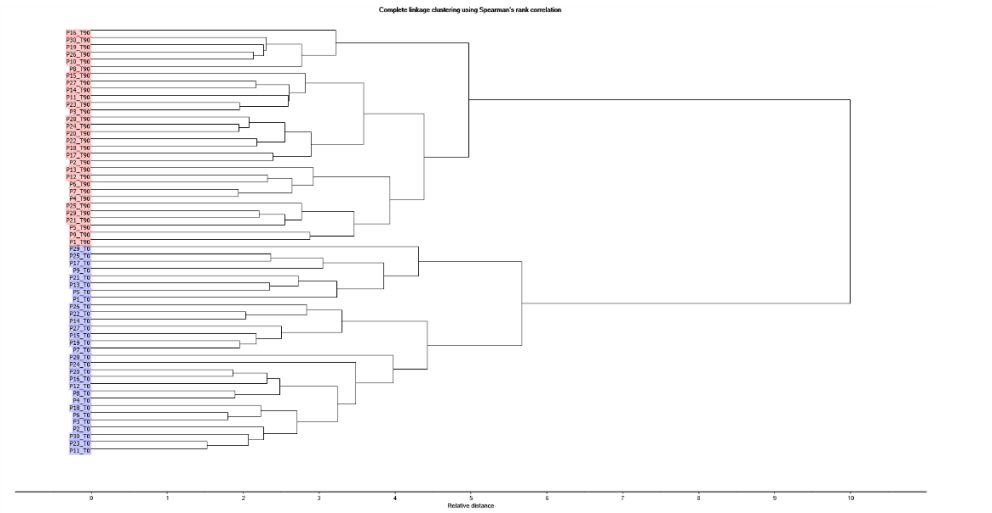


Figure 4.3.3. Hierarchical Cluster Analysis for the serum spectra, T0 and T90, pre-processed with atmospheric correction and a second derivative applied, with a second order polynomial, 15 smoothing points and a Savitzky-Golay filter, with HCA complete linkage and Spearman's rank correlation distance, without patient 10 on T0 to illustrate its effect on the HCA.

Interestingly, the hemoglobin of patient 10 at T0 was the lowest of all 30 volunteers, with a value approximately 20% lower than the average of the other 29 patients. It is also worth mentioning that whatever health imbalance the individual might have had, at the end of the study the hemoglobin level was normal. This could result from the ingestion of the extract since and as mentioned before, prolonged green tea consumption can, over time, reduce hemoglobin levels due to the antioxidants present in tea, which are known to hinder regular iron absorption [278], [279].

Table 4.3.1 summarizes the results obtained by HCA to discriminate T0 from T90 on plasma or serum samples. The combinations of pre-processing methods, the type of data (all replicas or the average of replicas) and the Kendall's Tau distance or the Spearman's rank correlation were evaluated. Out of all the combinations tested out, the atmospheric correction followed by the second derivative, allowed for a perfect separation of classes, with the sole exception of serum samples for the reduced data, where one patient did not fit into the classification algorithm.

Table 4.3.1. Representation by colors of HCA of FTIR spectra of plasma and serum for classification T0 from T90 samples: In green color, we refer to perfect separations of T0 and T90. In red the separation is not perfect and in orange the separation is incomplete leaving one patient (patient 10, T0), in a group of its own in the reduced data. It were used all the replicated samples or the reduced data set of the average of replicates. Were considered different spectral pre-processing methods and for HCA the Kendall's Tau distance or the Spearman's rank correlation was used.

		PLASMA		**SERUM**	
		distance measure			
	Pre-processing method	**Kendall's Tau**	**Spearman**	**Kendall's Tau**	**Spearman**
Replicates	*no pre-processing*				
	atmospheric + baseline correction				
	Atmospheric correction + second derivative				
Reduced = average of replicas	*no pre-processing*				
	atmospheric + baseline correction				
	Atmospheric correction + second derivative				

4.4. Regression methods

The main objective of regression models, as PCR and PLSR, is building a relationship between two groups of variables, the independent and the dependent variables. After a model has been constructed, it can be used for two separate actions:

- to *describe* said relationship between the two groups of variables, i.e., characterize our data and separate into clusters;

- to *predict* new values, i.e., upon existence of unlabeled samples, test the model and its capability of being able to predict where unknown samples would end up clustered.

In PCR and PLSR, the independent variables are called *predictors* (X-variable) and our dependent variables are designated as *responses* (Y-variables). With the present data, the X-variables are the spectra and the Y-variables are the constituents of the analyzed sample. In PLRS, the sum of squares of the deviations between the measured and predicted responses are minimized. To evaluate the regression model capacity the following parameters will be considered: *R-square*, *RMSE* and *Slope*. All spectra data considered the average of the replicate spectra. No outliers were removed.

PCR is a two-step procedure in which an X-matrix is first decomposed by PCA and is then fitted and MLR model, using the PC scores instead of the original X-variables as predictors. Since the scores are orthogonal, the PCR model does not suffer from collinearity effects, which is why some data analysts prefer PCR over PLS as it forces the user to better understand their data and its transformations, before finally applying the regression procedure.

In Figure 4.4.1, it was observed very similar PCR and PLSR scores based on plasma and serum data. Very good regression models were developed either with PCR and PLSR for either plasma or serum as pointed by Figure 4.4.2 and Figure 4.4.3 as well as Table 4.4.1 through Table 4.4.4. However, PLS did perform slightly better, with higher values of R-square and lower of RMSE obtained, and based on lesser factors (3), than those used for PCR (7 components), making it also faster to compute than the PCR counterpart.

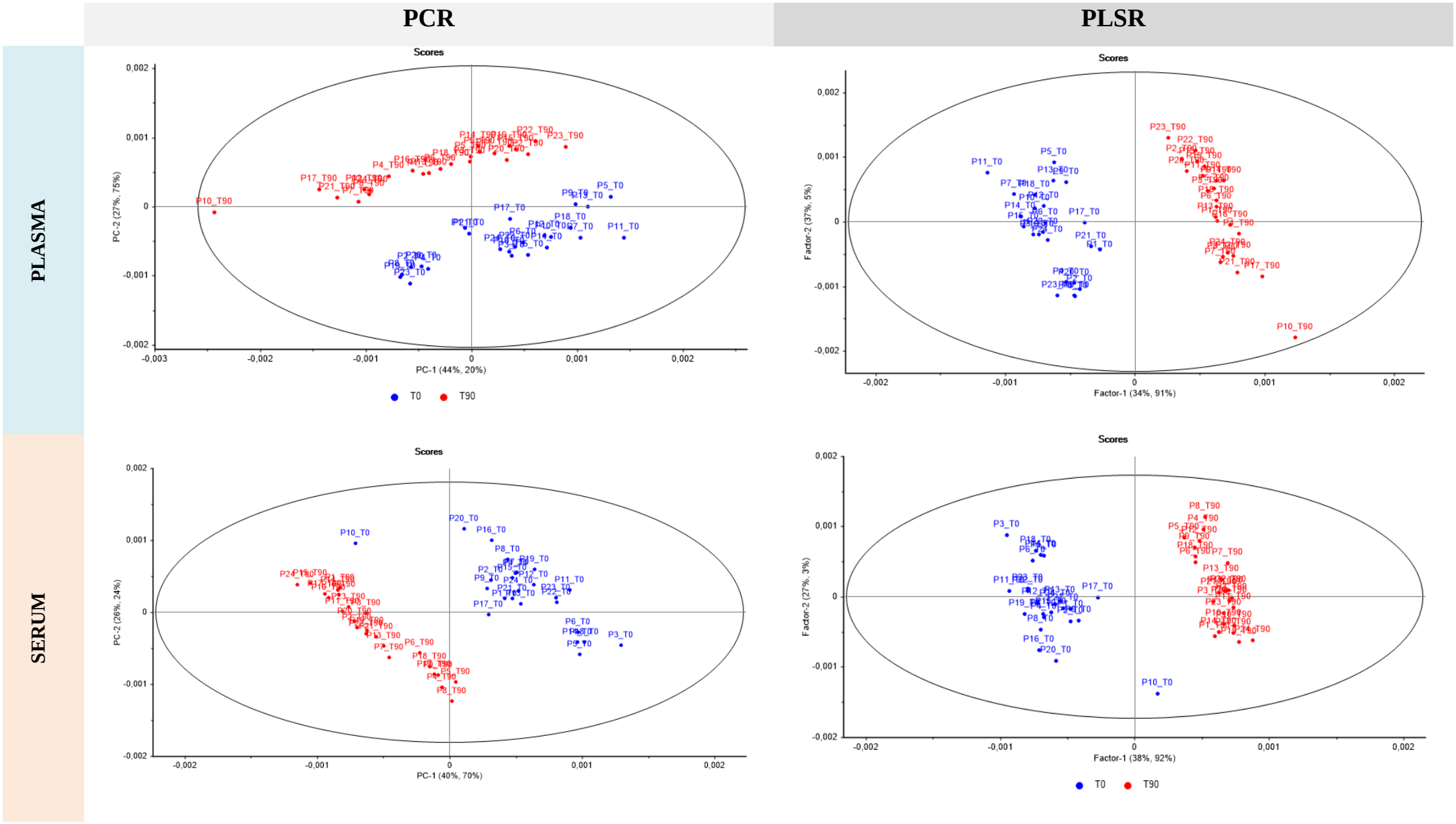


Figure 4.4.1. PCR and PLSR scores diagram with Hotelling's T^2 ellipse at 1% significance for spectra from Plasma and Serum at T0 or T90, and pre-processed with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15-points window. The number of PCs in PCR are 7 and the number of factors in PLSR are 3.

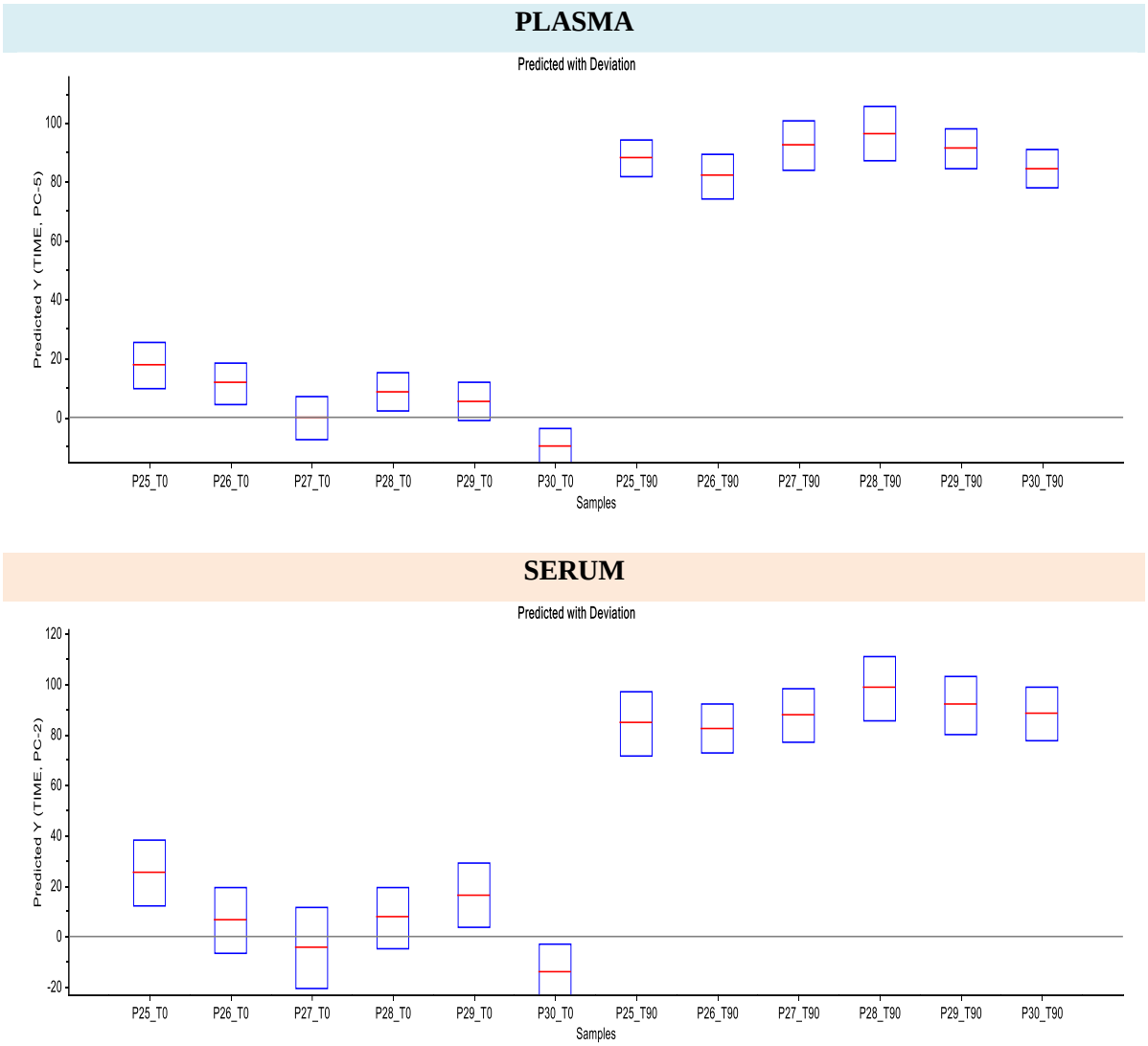


Figure 4.4.2. Predicted with deviation values for samples used for validation for PCR regression model for spectra of plasma and serum.

Table 4.4.1. Predicted Y-values and corresponding deviation values for PCR model with PC7 for plasma and serum.

		T0						T90					
		P25	**P26**	**P27**	**P28**	**P29**	**P30**	**P25**	**P26**	**P27**	**P28**	**P29**	**P30**
PLASMA	**Predicted**	15,8	8.2	-0.9	10.2	3.3	-9.9	88.0	80.5	92.4	94.1	86.6	87.1
	Deviation	7.2	6.1	6.7	5.9	5.2	6.0	5.8	6.7	7.8	8.1	5.6	5.5
SERUM	**Predicted**	25.0	6.0	-4.8	7.1	16.0	-14.4	84.1	81.9	87.4	97.9	91.2	87.7
	Deviation	13.0	12.8	16.1	12.0	12.8	11.4	12.7	9.6	10.6	12.5	11.6	10.5

Table 4.4.2. PCR fit parameters in PC7 (calibration and validation) for spectra of plasma and serum.

		Slope	**RMSE**	**R-squared**
PLASMA	**Calibration**	0.98	5.19	0.98
	Prediction	0.96	6.70	0.97
SERUM	**Calibration**	0.96	7.81	0.96
	Prediction	0.94	10.42	0.94

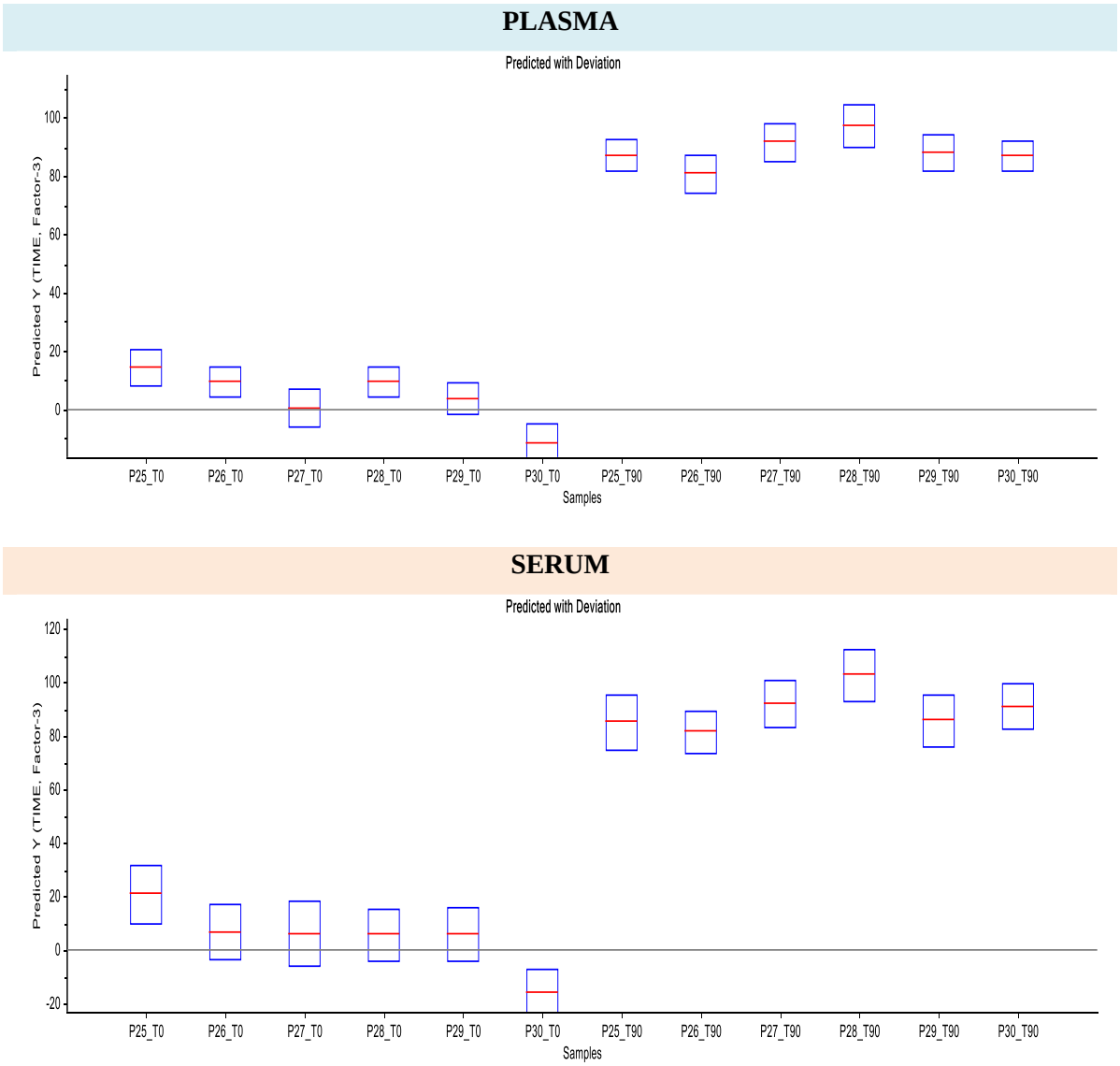


Figure 4.4.3. Predicted with deviation values for samples used for validation for PLSR regression model for spectra of plasma and serum.

Table 4.4.3. Predicted Y-values and corresponding deviation for PSLR with Factor 3 for spectra of plasma and serum.

		T0						T90					
		P25	**P26**	**P27**	**P28**	**P29**	**P30**	**P25**	**P26**	**P27**	**P28**	**P29**	**P30**
PLASMA	**Predicted**	14.1	9.3	0.3	9.1	3.4	-11.5	87.0	80.9	91.6	97.2	87.9	87.0
	Deviation	6.3	5.3	6.7	5.2	5.4	6.7	5.4	6.4	6.5	7.1	6.2	5.0
SERUM	**Predicted**	20.9	6.8	6.2	5.8	6.0	-16.0	85.3	81.4	92.1	102	85.8	91.0
	Deviation	10.7	10.3	12.0	9.6	9.8	9.2	10.3	7.8	8.6	9.9	9.5	8.4

Table 4.4.4. PLSR fit parameters in Factor 3 (calibration and prediction) for spectra of plasma and serum.

		Slope	**RMSE**	**R-squared**
PLASMA	**Calibration**	0.99	4.21	0.99
	Prediction	0.96	6.16	0.98
SERUM	**Calibration**	0.97	6.40	0.97
	Prediction	0.94	9.48	0.95

4.5. Discriminant Analysis

In this supervised classification method, classification rules are built upon prespecified classes (T0 and T90). These rules, that construct the model, are then used to allocate new and unknown samples to its most probable class. Discriminant Analysis (DA), is a type of qualitative calibration, where a category group variable is used for the classification instead of a continuous measurement, as in quantitative calibration. In this work DA was applied for the spectral data of plasma and serum.

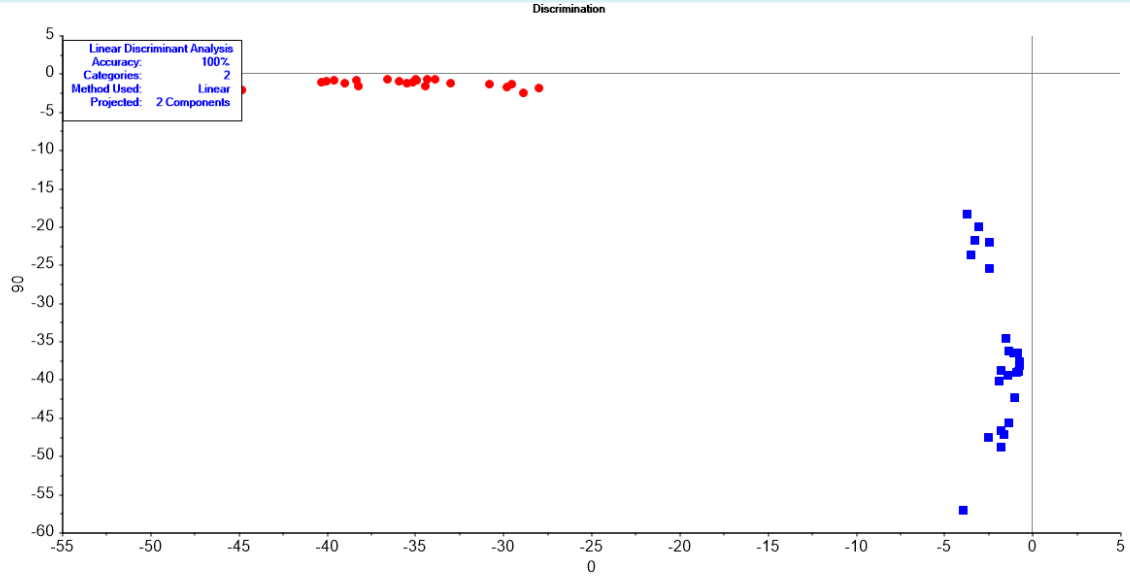
One of the techniques involved using a Linear Discriminant Analysis (LDA) in conjunction with the already available PCAs, leading to a PCA-LDA analysis. The linear method was chosen in detriment of a quadratic discriminant (used in cases where groups might have their main variability in different directions, and where a curve might best separate groups) or a Mahalanobis classifier (measure of distance between a point P and a distribution D), as it was found that the difference between the groups (T0 and T90) was best represented by a simple linear function. The second discriminant analysis method used was done by partial least squares regression methods (PLS-DA). All data in this chapter is regarding to spectra pre-processed by atmospheric correction followed by second derivative. All spectra data considered the average of the replicate spectra. No outliers were removed.

PCA-LDA

LDA plots of plasma and serum spectra between T0 and T90 and based on training data are represented in Figure 4.5.1. It was observed general good calibration models, as the samples categories, i.e. T0 and T90, either in the plasma or serum model, are both close to zero. Indeed, the confusion matrix, relative to calibration data (Table 4.5.1), shows that the LDA model for plasma, resulted in all samples being well classified, whereas the LDA model with serum data presents one sample that was not well classified, patient 10 at T0, in accordance with its previous classification as an outlier. This sample was off its actual class by a mere 1.087 points. This can also be seen better in Figure 4.5.1, where the sample is highlighted with a green circle.

Concerning the prediction capacity, in both LDA models, either for plasma and serum (Table 4.5.1), all samples were predicted in its real class, i.e. T0 or T90.

PLASMA



SERUM

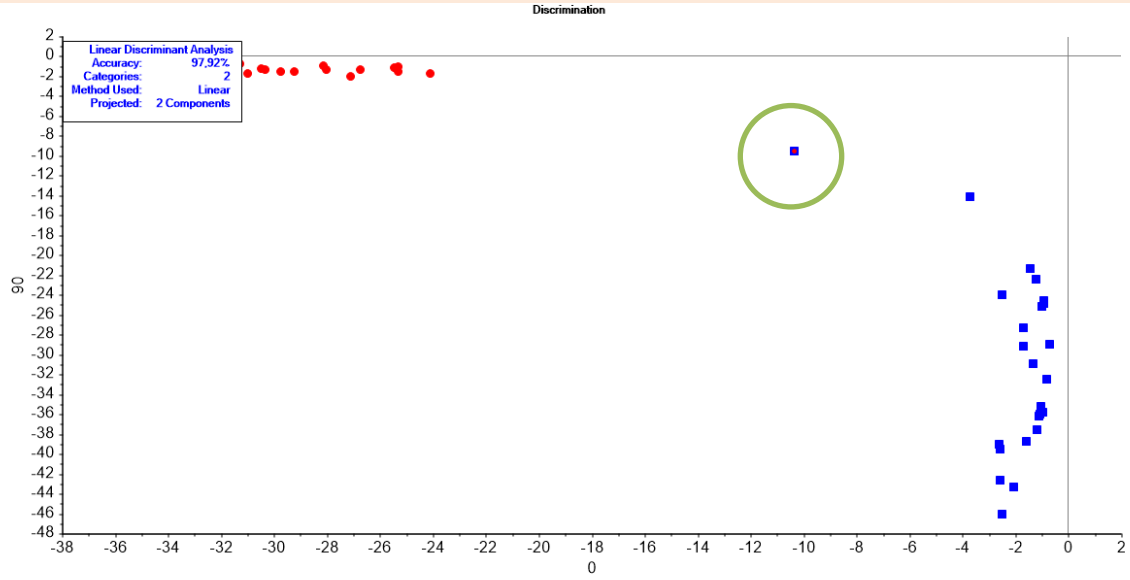


Figure 4.5.1. PCA-LDA discrimination plot for spectra of plasma and serum pre-processed with atmospheric correction and a second derivative, with a 2nd order polynomial and a Savitzky-Golay filter with 15-points window. Both LDA models had two projected components and achieved a 100% accuracy in the separation of both classes: T0 (in blue) and T90 (in red). These samples for both plasma and serum pertain to the calibration samples (first 24 volunteers out of the 30, for both T0 and T90).

Table 4.5.1. Confusion matrix of PCA-LDA for spectra of plasma or serum relative to T0 and T90. The data pertains to the calibration samples (first 24 volunteers out of the 30, for both T0 and T90).

PLASMA	Real	
Predicted	**0**	**90**
0	24	0
90	0	24

SERUM	Real	
Predicted	**0**	**90**
0	23	0
90	1	24

Table 4.5.2. Prediction matrix of PCA-LDA for spectra of plasma and serum relative to T0 and T90. The data pertains to the validation (test) samples (last 6 volunteers out of 30, for both T0 and T90).

	Test Samples	Real class	T0	T90	Predicted class
PLASMA	P25	0	-3,65874	-19,8294	0
	P26	0	-0,87238	-33,4946	0
	P27	0	-3,36185	-48,2916	0
	P28	0	-2,69111	-39,6309	0
	P29	0	-2,28761	-25,4981	0
	P30	0	-2,10727	-43,705	0
	P25	90	-39,2311	-1,43648	90
	P26	90	-29,0563	-1,94352	90
	P27	90	-34,1022	-4,07935	90
	P28	90	-41,906	-1,20278	90
	P29	90	-39,6587	-0,80424	90
	P30	90	-30,6125	-0,96566	90
SERUM	P25	0	-2,72596	-16,4292	0
	P26	0	-3,11375	-29,8764	0
	P27	0	-1,88341	-36,1394	0
	P28	0	-1,08175	-27,1093	0
	P29	0	-1,88522	-21,7878	0
	P30	0	-2,33948	-43,2029	0
	P25	90	-28,8195	-1,91709	90
	P26	90	-27,1148	-1,6954	90
	P27	90	-29,9583	-0,7596	90
	P28	90	-37,5687	-1,16218	90
	P29	90	-33,1743	-1,33236	90
	P30	90	-30,159	-0,76858	90

PLS-DA

PLS-DA uses PLS regression for discrimination or classification purposes and it is based on the modelling of the differences between the two classes, T0 and T90. The PLS-DA model used the following class memberships: -1 for members of class T0 and +1 for members of class T90.

It was observed that either the PLS-DA model for plasma or the one for serum (Table 4.5.3), predicted correctly all 12 samples, as samples from real class T0 presented values near -1, were samples from real class T90 presented values near +1.

Table 4.5.3. PLS-DA prediction classes for plasma and serum test samples.

PLASMA			SERUM		
Test samples	**Predicted**	**Deviation**	**Test samples**	**Predicted**	**Deviation**
P25	-0,892873	0,1063112	P25	-0,8492607	0,1858644
P26	-1,005116	0,07649239	P26	-1,082608	0,2018854
P27	-1,0183	0,1230721	P27	-1,046612	0,2486495
P28	-0,9690454	0,1232325	P28	-1,039275	0,2095353
P29	-0,9355171	0,09070035	P29	-1,042276	0,2271922
P30	-1,050511	0,1066955	P30	-1,077311	0,1787859
P25	1,015758	0,08775617	P25	0,9593408	0,2110093
P26	0,9784197	0,09249701	P26	0,8480738	0,1681136
P27	0,9933492	0,149099	P27	1,026433	0,1909677
P28	1,050914	0,09667311	P28	0,9479635	0,1852195
P29	0,8709764	0,07453051	P29	0,7881708	0,1744379
P30	0,9572209	0,07924302	P30	0,9470502	0,1804974

4.6. Uni-variate Spectral Analysis

This subchapter was divided into two parts:

- In the first one, the differences between plasma and serum spectra were analyzed;
- In the second part, a statistical analysis of diverse ratios of spectral peaks, either for plasma or serum data, between T0 and T90, were analyzed.

Differences between plasma and serum spectra

The averaged spectra of plasma of the 30 participants pre-processed with atmospheric and baseline correction were compared with the average spectra of serum (Figure 4.6.1), either at T0 or T90. It was observed that the average spectra of spectrum present slightly higher absorbance values than the average spectra of plasma (Figure 4.6.1), in accordance with other authors work [280]. This could seem to go against what was expected, since serum does not contain fibrinogen in contrast to plasma. However, serum presents a higher diversity and concentration of metabolites when compared to plasma [129]. Interestingly, in both serum and plasma spectrum, the major absorbance peaks (Amide I and II, approximately at 1650 and 1550cm^{-1}, respectively), in T0, seem to present marginally higher values than in T90 (Figure 4.6.1), highlighting the effect of EGCG in these biofluids molecular profile.

The second derivative spectra of plasma were compared against the second derivative of serum at T0 and T90, respectively (Figure 4.6.2). The second derivative spectra were obtained from the average spectra of the 30 participants pre-processed by atmospheric correction. The second derivative spectra highlight the chemical differences between these two biofluids. For simplicity, Figure 4.6.2 highlights the spectral regions between 2000 and 1000 cm^{-1}.

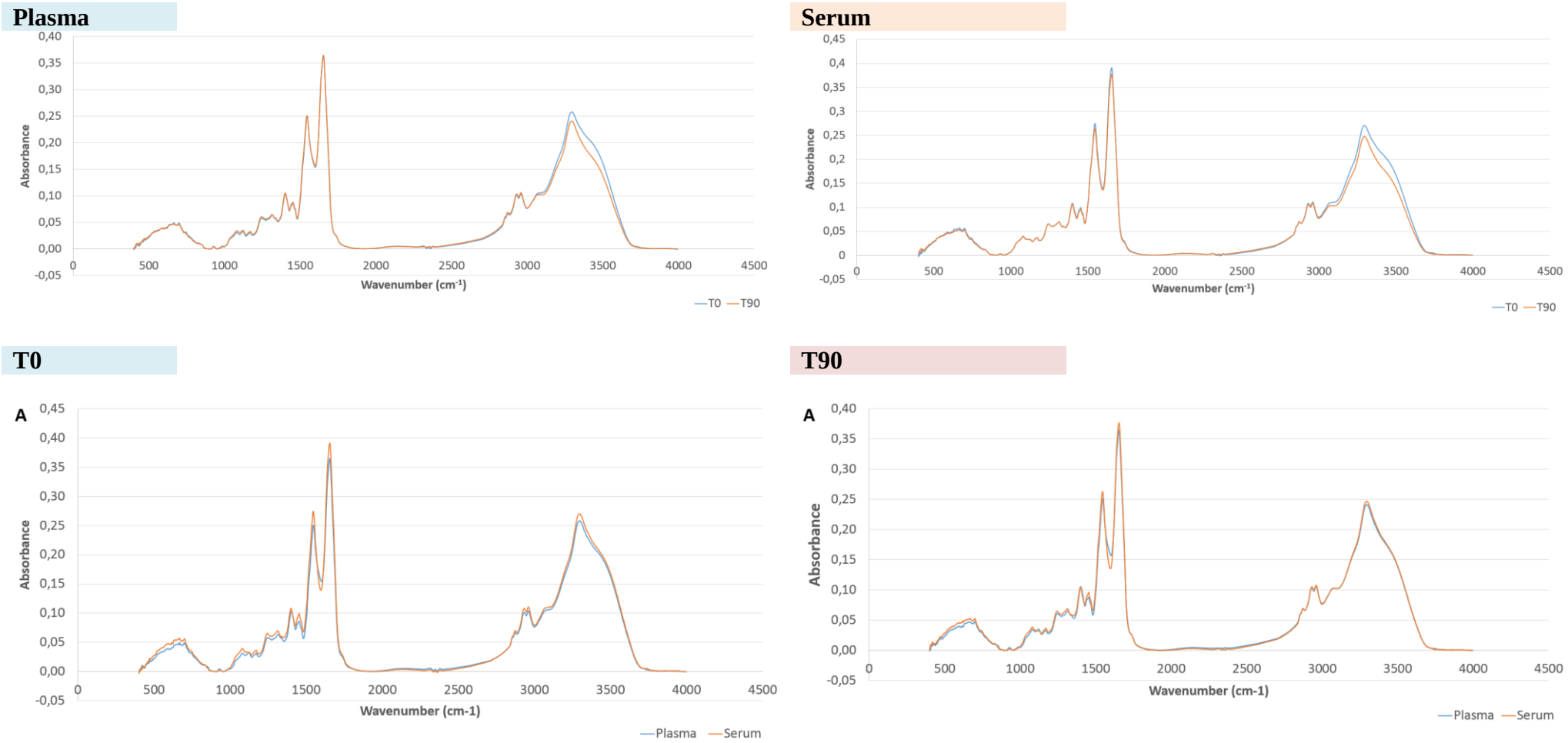


Figure 4.6.1. Above: average spectra of the 30 participants for the preprocessed spectra of plasma and serum for atmospheric correction and baseline correction, with T0 (blue) and at T90 (red). Bellow: average spectra of the 30 participants for the pre-processed spectra of plasma (blue) and serum (orange), for atmospheric correction and baseline correction, with T0 (left) and at T90 (right).

Figure 4.6.2. Second derivative of the average spectra (pre-processed by atmospheric correction) of the 30 participants at T0 (above) and T90 (below) for plasma (blue) and serum (orange). 2nd derivative was based on a Savitzky-Golay filter with 2nd order polynomial with a 15-points window.

To quantify the different molecular profiles between plasma and serum, the negative peaks of the second derivative were considered, which correspond to the positive peaks in the normal spectra, at T0 (Table 4.6.1). The absorbance peaks were represented by the "A" letter, followed by the number of the corresponding wavenumber. For example, A1550, regards the band obtained at the wavenumber of 1550 cm^{-1}. The difference in percentage between serum and plasma is presented in Table 4.6.1. On it, positive values correspond to serum presenting absorbance values higher than plasma and *vice versa*. Many regions of the second derivative spectra were different between plasma and serum, either at T0 or T90. Therefore, a highly different molecular profile was observed between plasma and serum, as expected.

Table 4.6.1. Percentage of the differences between negative peaks of the second derivative spectra of serum in relation to plasma at T0 and T90, respectively, as represented in figure 4.6.2.

WAVENUMBERS (cm^{-1})	% difference between serum and plasma	WAVENUMBERS (cm^{-1})	% difference between serum and plasma
3991,341	28%	**2006,76**	58%
3973,019	-52%	**1967,223**	-77%
3946,018	24%	**1799,43**	100%
3916,124	31%	**1779,18**	-19%
3869,836	80%	**1766,644**	695%
3831,263	-30%	**1691,426**	36%
3809,084	16%	**1679,854**	33%
3784,976	-65%	**1657,675**	-1%
3744,474	1557%	**1639,353**	15%
3725,188	-41%	**1621,031**	-7%
3664,435	-986%	**1615,244**	-9%
3567,038	-26%	**1583,422**	1108%
3555,466	-92%	**1545,813**	7%
3535,216	-97%	**1534,241**	80%
3524,608	-7%	**1516,883**	18%
3513,036	52%	**1498,561**	-46%
3500,5	-42%	**1469,631**	12%
3345,244	-131%	**1439,737**	20%
3299,92	1%	**1400,2**	19%
3265,205	7%	**1243,015**	29%
3240,132	-45%	**1189,013**	-21%
3206,381	-39%	**1172,619**	9%
3172,629	-59%	**1154,297**	10%
3145,628	43%	**1128,26**	-20%
3083,911	-29%	**1030,864**	18%
2925,762	8%	**973,9683**	-225%
2871,76	22%	**970,1111**	-174%
2794,614	-308%	**951,7888**	-53%
2636,465	-97%	**898,751**	-50%
2579,569	-99%	**881,3931**	488%
2483,137	-192%	**875,6072**	-76%
2283,522	-1%	**787,8535**	-16%
2201,554	-45%	**729,0298**	-44%
2135,98	27%	**572,8091**	-5%
2051,119	124%	**519,7712**	182%
2031,833	-179%	**426,2317**	-13%

Ratios of spectral peaks to discriminate T90 from T0

The 2nd derivative spectra with atmospheric correction of the average of all spectra of plasma (Figure 4.6.3) and serum (Figure 4.6.4) at T0 and at T90, were considered. Only the negative peaks were considered, as they represent positive peaks in a normal spectrum. The spectral bands that presented differences between T0 and T90 on plasma and serum were registered on Table 4.6.2. This information was used to determine ratios of spectral peaks on the second derivative spectra of the 30 participants. Ratios were considered, instead of the absorbance of one unique peak, to minimize the effect of sample quantity, while highlighting the chemical information on the spectra. The ratios were determined between the peaks (identified in Table 4.6.3) and reference peaks were chosen as the ones presenting almost no changes between in absorbance value between T0 and T90. Two reference peaks were considered, one for the region between 400 to 1800 cm^{-1} and another for the 2800 to 4000 cm^{-1} regions. These reference peaks are usually prominent in human plasma [281], [282], [283] and serum [280] [284] spectra. These reference bands were 2924 and 962 cm^{-1} for plasma and 2871 and 1172 cm^{-1} for serum. The average and standard deviation of the ratios between peaks for plasma and serum are represented in Table 4.6.3 and Table 4.6.4, respectively. Boxplots of some of these ratios for plasma and serum are represented in Figure 4.6.5 and Figure 4.6.6, respectively.

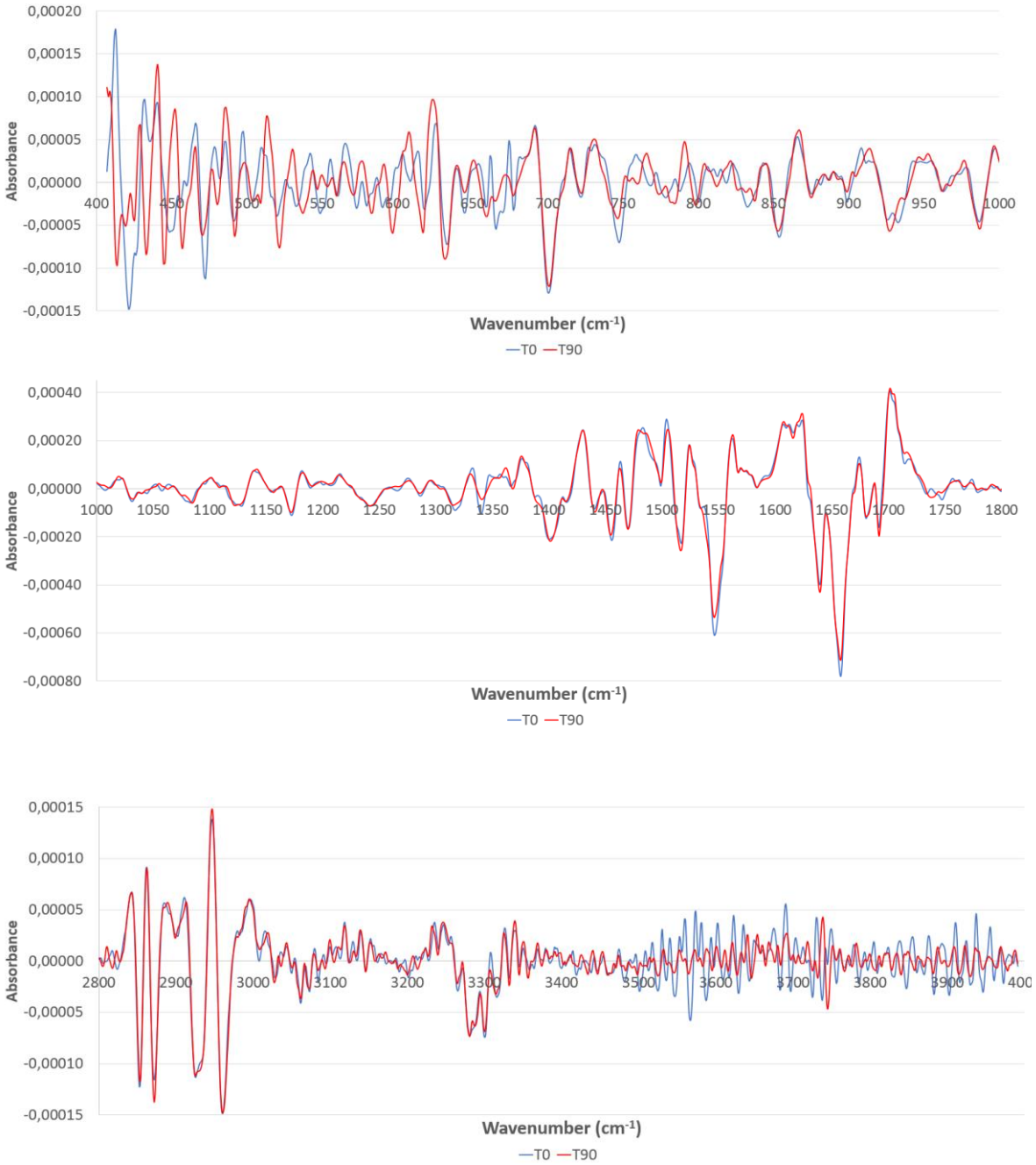


Figure 4.6.3. Second-derivative of the averaged spectra of plasma of the 30 participants, at T0 group (blue line) and T90 (red line).

Figure 4.6.4. Second-derivative of the averaged spectra of serum of the 30 participants, at T0 group (blue line) and T90 (red line).

Table 4.6.2. Percentage of the differences between negative peaks of the second derivative spectra at T0 and T90, from plasma and serum samples, respectively, and as represented in Figures 4.6.3 and 4.6.4. A total of 48 spectral peaks in plasma spectra and 54 in serum spectra were identified as different between T0 and T90.

PLASMA			
Spectral band	**(T90 vs T0 - %)**	**Spectral band**	**(T90 vs T0 - %)**
3991	-57	1679	-5
3946	-73	1657	-9
3929	-35	1639	8
3869	-58	1615	-11
3814	1	1583	-37
3766	59	1567	-6
3756	-104	1545	-12
3744	1358	1498	94
3725	-73	1469	3
3697	-103	986	17
3664	-414	962	-1
3657	0	898	-51
3524	-50	853	-11
3491	256	722	-24
3463	2	676	-51
3446	30	665	-61
3345	155	491	40
3299	-8		
3280	1		
3265	-25		
3213	-737		
3145	18		
3083	-56		
3061	-11		
2960	-1		
2924	-1		
2871	18		
2805	164		
1779	-63		
1766	-358		
1691	23		

SERUM			
Spectral band	**(T90 vs T0 - %)**	**Spectral band**	**(T90 vs T0 - %)**
3991	113	1734	-57
3932	-60	1691	-10
3916	-96	1657	-8
3900	-60	1594	-17
3822	-687	1545	-11
3744	-117	1516	5
3738	-249	1498	232
3645	-4	1469	-4
3595	-28	1439	-8
3585	-7	1418	-637
3576	-19	1400	-6
3535	2068	1348	53
3524	-19	1340	5
3482	1-6	1289	30
3374	-208	1172	-4
3355	-35	1012	160
3345	18	973	-87
3312	-3	898	16
3287	-10	881	61
3278	10	833	2
3240	-5	787	-81
3226	-164	519	-47
3219	-427	506	-70
3212	-122		
3199	-5		
3159	-249		
3093	49		
3060	9		
2871	-2		
2853	-15		
1779	-57		

Table 4.6.3. Ratios of spectral peaks for plasma and serum.

Peak Ratios for plasma data	A3991/A2924	A3946/A2924	A3929/A2924	A3869/A2924	A3814/A2924
	A3766/A2924	A3756/A2924	A3744/A2924	A3725/A2924	A3697/A2924
	A3664/A2924	A3657/A2924	A3524/A2924	A3491/A2924	A3463/A2924
	A3446/A2924	A3345/A2924	A3299/A2924	A3280/A2924	A3265/A2924
	A3213/A2924	A3145/A2924	A3083/A2924	A3061/A2924	A2960/A2924
	A2871/A2924	A2805/A2924	A1779/A962	A1766/A962	A1691/A962
	A1679/A962	A1657/A962	A1639/A962	A1615/A962	A1583/A962
	A1567/A962	A1545/A962	A1498/A962	A1469/A962	A986/A962
	A898/A962	A853/A962	A722/A962	A676/A962	A665/A962
	A491/A962				

Peak Ratios for serum data	A3991/A2871	A3932/A2871	A3916/A2871	A3900/A2871	A3822/A2871
	A3744/A2871	A3738/A2871	A3645/A2871	A3595/A2871	A3585/A2871
	A3576/A2871	A3535/A2871	A3524/A2871	A3482/A2871	A3374/A2871
	A3355/A2871	A3345/A2871	A3312/A2871	A3287/A2871	A3278/A2871
	A3240/A2871	A3226/A2871	A3219/A2871	A3212/A2871	A3199/A2871
	A3159/A2871	A3093/A2871	A3060/A2871	A2853/A2871	A1179/A1172
	A1174/A1172	A1691/A1172	A1657/A1172	A1594/A1172	A1545/A1172
	A1516/A1172	A1498/A1172	A1469/A1172	A1439/A1172	A1418A1172
	A1400/A1172	A1348/A1172	A1340/A1172	A1289/A1172	A1012/A1172
	A973/A1172	A898/A1172	A881/A1172	A833/A1172	A787/A1172
	A519/A1172	A506/A1172			

A *t*-student test was applied to compare the ratios between spectral peaks between T0 and T90. Table 4.6.4 and Table 4.6.5, presents the *p*-values of the *t*-student test for these ratios for plasma and serum spectra.

Regarding plasma, out of the 46 peak ratios determined, 16 presented values statistically different between the group that ingested the EGCG extract after 90 days versus the same group at T0 (before ingestion) at 1% significance. These include bands associated to lipids and proteins as A2871/A2824, A2805/A2924, A3213/A2924 and A3083/A2924.

For serum, a total of 52 peak ratios were determined and from these, 33 presented values statistically different between T90 and T0 at 1% significance, with bands associated to lipids, proteins and RNA: A3535/A2871, A3374/A2871, A3355/A2871, A3219/A2871, A3060/A2871, A1545/A1172 and A2853/A2871.

The difference in the number of peak ratios statistically different between T0 and T90, observed in serum (n=33) in relation to the number of peak ratios statistically different observed with plasma data (n=16), is more than double, pointing out serum as a richer biofluid in molecular information in the analysis of the effect of EGCG consumption. This observation is in accordance to serum presenting a significantly higher number of metabolites than plasma [129].

*Table 4.6.4. Average values and standard deviations of spectral absorbance ratios of human **plasma** diluted at 1/10 for groups T0 and T90 and p-value of student's t-test regarding the comparison of spectral bands of T0 and T90 group.*

Spectral ratios	**T0 group** *(before EGCG ingestion)*		**T90 group** *(90 days after EGCG ingestion)*		***p*-value** of *t*-student analysis (α<0,01)
	Average	Standard deviation	Average	Standard deviation	
A3991/A2924	0,046	0,047	0,022	0,044	0,046
A3946/A2924	0,286	0,145	0,080	0,044	0,000
A3929/A2924	0,219	0,077	0,143	0,046	0,001
A3869/A2924	0,121	0,075	0,054	0,062	0,000
A3814/A2924	0,046	0,096	0,053	0,083	0,771
A3766/A2924	0,060	0,050	0,100	0,077	0,041
A3756/A2924	0,202	0,090	-0,010	0,080	0,000
A3744/A2924	0,033	0,167	0,436	0,121	0,000
A3725/A2924	0,339	0,222	0,090	0,068	0,000
A3697/A2924	0,283	0,155	-0,009	0,051	0,000
A3664/A2924	-0,013	0,054	0,045	0,077	0,003
A3657/A2924	-0,066	0,072	-0,074	0,065	0,663
A3524/A2924	0,273	0,118	0,138	0,071	0,000
A3491/A2924	0,015	0,040	0,063	0,045	0,000
A3463/A2924	0,108	0,043	0,113	0,049	0,696
A3446/A2924	0,093	0,062	0,125	0,058	0,078
A3345/A2924	0,050	0,051	0,123	0,043	0,000
A3299/A2924	0,684	0,156	0,633	0,111	0,133
A3280/A2924	0,666	0,141	0,686	0,127	0,594
A3265/A2924	0,269	0,068	0,206	0,064	0,001
A3213/A2924	-0,015	0,036	0,077	0,044	0,000
A3145/A2924	0,082	0,040	0,099	0,035	0,092
A3083/A2924	0,192	0,051	0,086	0,033	0,000
A3061/A2924	0,376	0,078	0,338	0,063	0,030
A2960/A2924	1,352	0,255	1,346	0,194	0,907
A2871/A2924	1,053	0,232	1,256	0,190	0,000
A2805/A2924	0,017	0,019	0,050	0,032	0,000
A1779/A962	-1,382	16,526	0,824	1,006	0,479
A1766/A962	10,025	51,722	-1,205	1,468	0,251
A1691/A962	-154,421	926,927	29,074	20,961	0,297
A1679/A962	-118,618	710,960	15,580	10,049	0,317
A1657/A962	-1074,811	6243,124	108,797	94,811	0,317
A1639/A962	-402,992	2408,620	70,511	62,265	0,303
A1615/A962	265,936	1570,306	-31,814	26,056	0,317
A1583/A962	11,973	68,204	-1,244	2,561	0,311
A1567/A962	125,249	714,031	-10,809	11,870	0,316
A1545/A962	-764,353	4474,337	81,826	68,728	0,319
A1498/A962	-4,277	15,662	-3,243	2,199	0,717
A1469/A962	-214,968	1253,398	24,979	19,217	0,312
A986/A962	-51,925	306,203	7,665	5,562	0,304
A898/A962	-28,994	169,524	1,855	2,066	0,336
A853/A962	-81,942	477,805	8,193	6,524	0,319
A722/A962	-7,973	53,104	2,406	3,025	0,322
A676/A962	-7,584	61,767	2,602	3,354	0,387
A665/A962	-55,517	331,836	3,848	6,940	0,346
A491/A962	-29,204	186,372	10,387	10,217	0,270

*Table 4.6.5. Average values and standard deviations of spectral absorbance ratios of human **serum** diluted at 1/10 for groups T0 and T90 and p-value of student's t-test regarding the comparison of spectral bands of T0 and T90 group.*

Spectral ratios	**T0 group** *(before EGCG ingestion)*		**T90 group** *(90 days after EGCG ingestion)*		***p*-value** of *t*-student analysis (α<0,01)
	Average	Standard deviation	Average	Standard deviation	
A3991/A2871	0,041332135	0,036309717	0,091711469	0,037516991	0,000
A3932/A2871	0,220170574	0,088766922	0,091310134	0,027671775	0,000
A3916/A2871	0,197655467	0,089037713	0,010188576	0,03234847	0,000
A3900/A2871	0,369632448	0,135991316	0,151614341	0,062450941	0,000
A3822/A2871	0,00632087	0,050878163	-0,027701841	0,047110159	0,006
A3744/A2871	0,447727328	0,141821095	-0,076716848	0,07246343	0,000
A3738/A2871	-0,121560494	0,048701476	0,185900967	0,123295665	0,000
A3645/A2871	0,236330898	0,132878618	0,233759268	0,109343957	0,934
A3595/A2871	0,067037341	0,05257882	0,049604216	0,047943505	0,188
A3585/A2871	0,18692218	0,113581884	0,179586135	0,054248471	0,727
A3576/A2871	-0,127270113	0,061567215	-0,102234966	0,045262066	0,062
A3535/A2871	0,000707842	0,038940531	0,05933428	0,028293971	0,000
A3524/A2871	0,194861526	0,103089143	0,160345201	0,048801396	0,041
A3482/A2871	0,137410282	0,049825853	0,121100844	0,043910827	0,214
A3374/A2871	-0,031085261	0,043999646	0,032418661	0,032013676	0,000
A3355/A2871	0,138147776	0,044213108	0,090778284	0,032275703	0,000
A3345/A2871	-0,007975834	0,045879892	-0,014430573	0,027998724	0,538
A3312/A2871	0,219888219	0,035428886	0,21904618	0,033877926	0,919
A3287/A2871	0,515943791	0,058768877	0,470768384	0,046496255	0,002
A3278/A2871	0,477442143	0,046564257	0,534511984	0,041423157	0,000
A3240/A2871	-0,045990057	0,03400681	-0,043368004	0,032303882	0,730
A3226/A2871	-0,044173594	0,033611411	0,031323729	0,031840691	0,000
A3219/A2871	0,014655423	0,048045853	-0,050335755	0,027417805	0,000
A3212/A2871	0,040095948	0,027977661	-0,008628674	0,027337954	0,000
A3199/A2871	0,079023249	0,034741946	0,076382465	0,024378513	0,757
A3159/A2871	-0,020301698	0,029780639	0,032600609	0,024235507	0,000
A3093/A2871	0,059475148	0,020703525	0,089798069	0,021956619	0,000
A3060/A2871	0,262532527	0,031403604	0,290151662	0,025299916	0,002
A2853/A2871	0,986376186	0,232701528	0,844167298	0,228815465	0,000
A1779/A1172	0,097812265	0,058364593	0,044613745	0,052800049	0,000
A1734/A1172	0,28644109	0,098704235	0,130862664	0,101580697	0,000
A1691/A1172	1,847186479	0,29677015	1,731477988	0,265767014	0,012
A1657/A1172	6,533223371	0,711675367	6,311211473	0,617073701	0,230
A1594/A1172	-1,278085929	0,107435949	-1,111965373	0,087170161	0,000
A1545/A1172	5,52241956	0,472186109	5,142150902	0,455327704	0,003
A1516/A1172	2,250525471	0,210627341	2,464845665	0,200293629	0,000
A1498/A1172	-0,050484441	0,106203706	-0,178893924	0,082722875	0,000
A1469/A1172	1,551364071	0,094247546	1,551969168	0,091393512	0,965
A1439/A1172	0,841300161	0,072304254	0,807143592	0,083371604	0,102
A1418/A1172	0,047553866	0,12311592	-0,244474351	0,059436584	0,000
A1400/A1172	2,102765221	0,098057135	2,06395717	0,105379878	0,059
A1348/A1172	-0,03753012	0,046767371	-0,062929683	0,049082098	0,002
A1340/A1172	0,315314506	0,152055239	0,350496857	0,111036658	0,150
A1289/A1172	0,03859354	0,035074642	0,051518881	0,030222874	0,070
A1012/A1172	0,051262579	0,034574491	0,141102552	0,029834031	0,000
A973/A1172	0,103681612	0,045070722	0,014581263	0,038476244	0,000
A898/A1172	0,096333165	0,047578037	0,117035285	0,037123814	0,082
A881/A1172	0,126301603	0,055567284	0,218146495	0,048884413	0,000
A833/A1172	0,428608306	0,082030096	0,455834919	0,077448039	0,252
A787/A1172	0,072118629	0,054225374	0,014286723	0,051944031	0,000
A519/A1172	0,954941427	0,193337044	0,52746486	0,160842421	0,000
A506/A1172	0,220200767	0,225121075	0,071671388	0,142423585	0,007

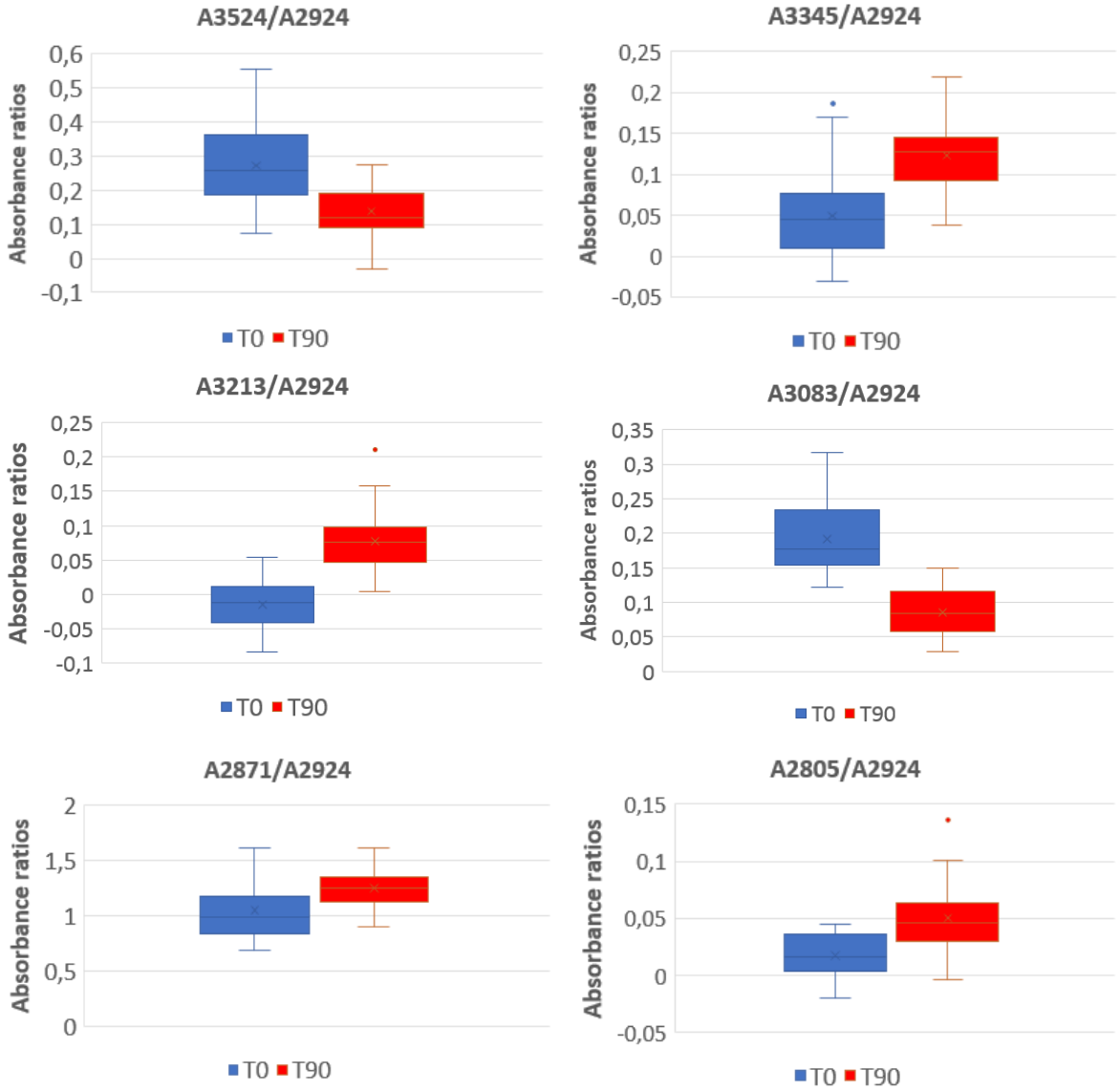


Figure 4.6.5. Boxplots for some of the absorbance ratios of the T0 group (blue) and T90 (red) for human plasma.

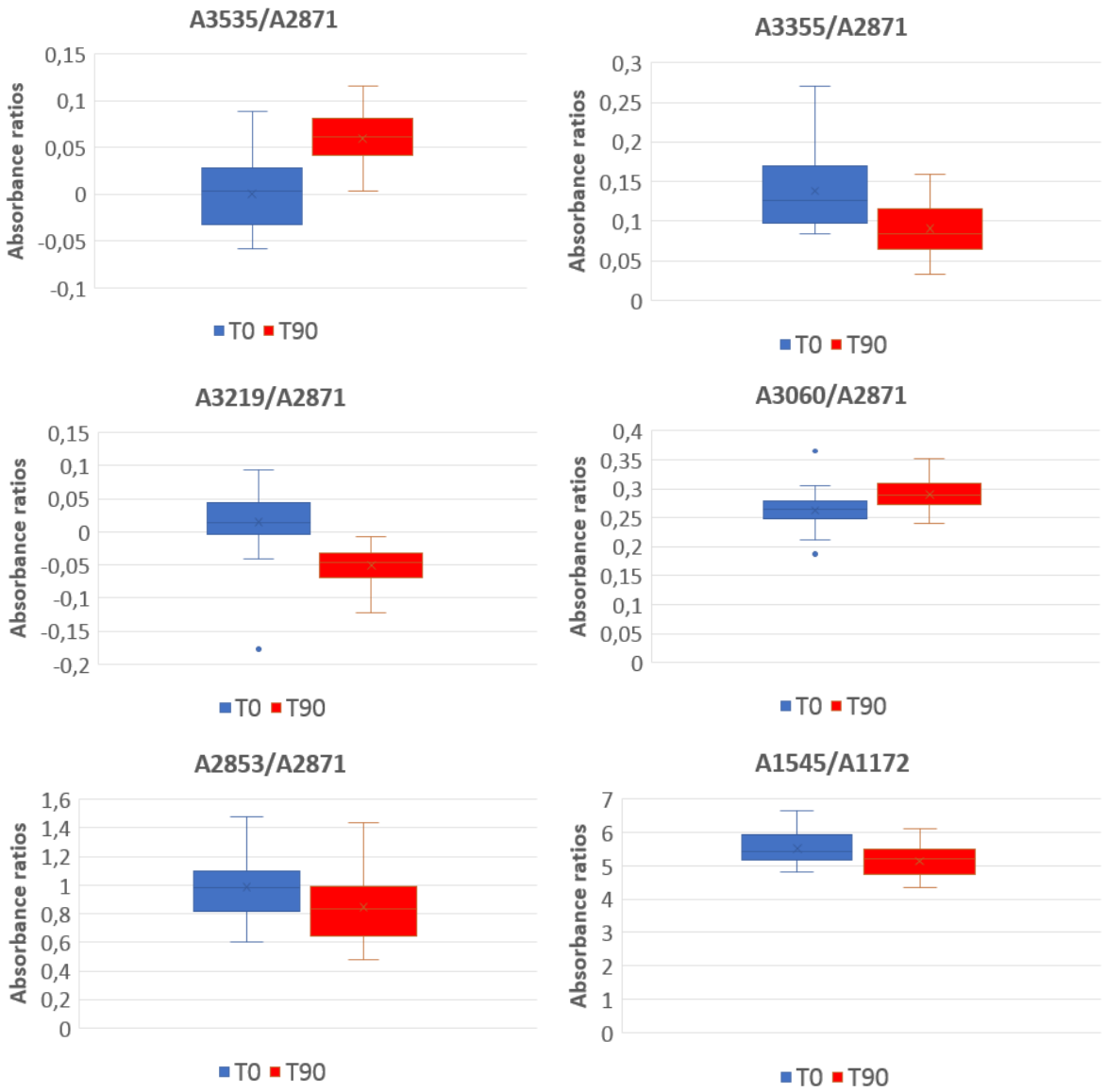


Figure 4.6.6. Boxplots for the some of the absorbance ratios of the T0 group (blue) and T90 (red) for human serum.

Chapter 5: Machine Learning Application

5.1. Aims

In this chapter, it was aimed to develop a platform, based on a rapid visual workflow, easy to use, automatic to analyze the pre-processing and processing methods of spectral data (as associated to biofluids as serum and plasma), based on Orange software. As such, it is expected that the platform enables to conduct, automatically, the associated pre-processing and processing analysis, as unsupervised and supervised methods (e.g., PCA, PCR, PLSR, HCA, etc.), search for correlations between principal components/latent variables, specify relevant spectral regions, capable of prediction capability of unlabeled samples and even be able to perform, with statistical significance, absorbance ratios.

5.2. Workflows

The following two main workflows were developed:

- The first, simpler workflow (Figure 5.2.1), evaluates the 7 blood tests analysis that were statistically different between T90 and T0, according to subchapter 4.1 and as described in Figure 5.2.2 through Figure 5.2.7;
- The second is the main and more complex workflow (Figure 5.2.8), that includes the spectra input of data to the end result.

Workflow to analyse clinical blood analytes

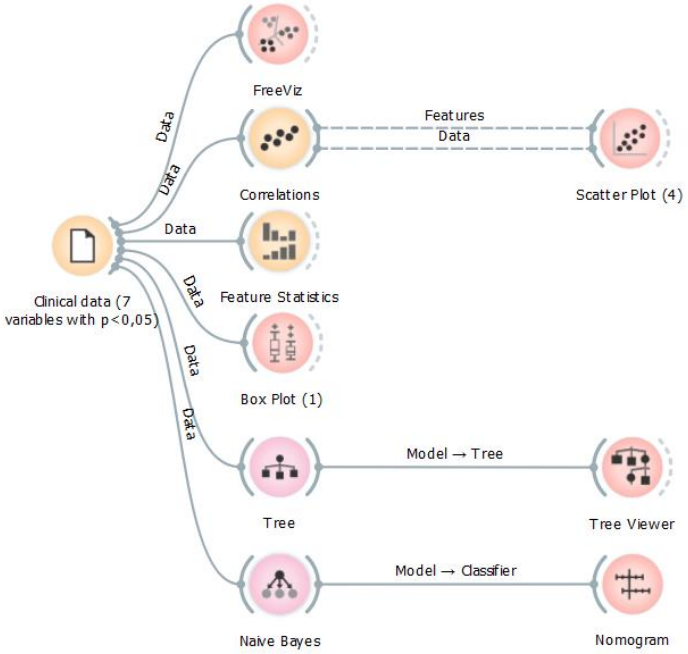


Figure 5.2.1. Workflow to analyze 7 blood clinical variables of the 30 patients.

The main tools/widgets used and outputs observed with the blood clinical analysis are briefly described next, along with their visual representations (blue is T0 and red is T90), in Figure 5.2.1 through Figure 5.2.7.

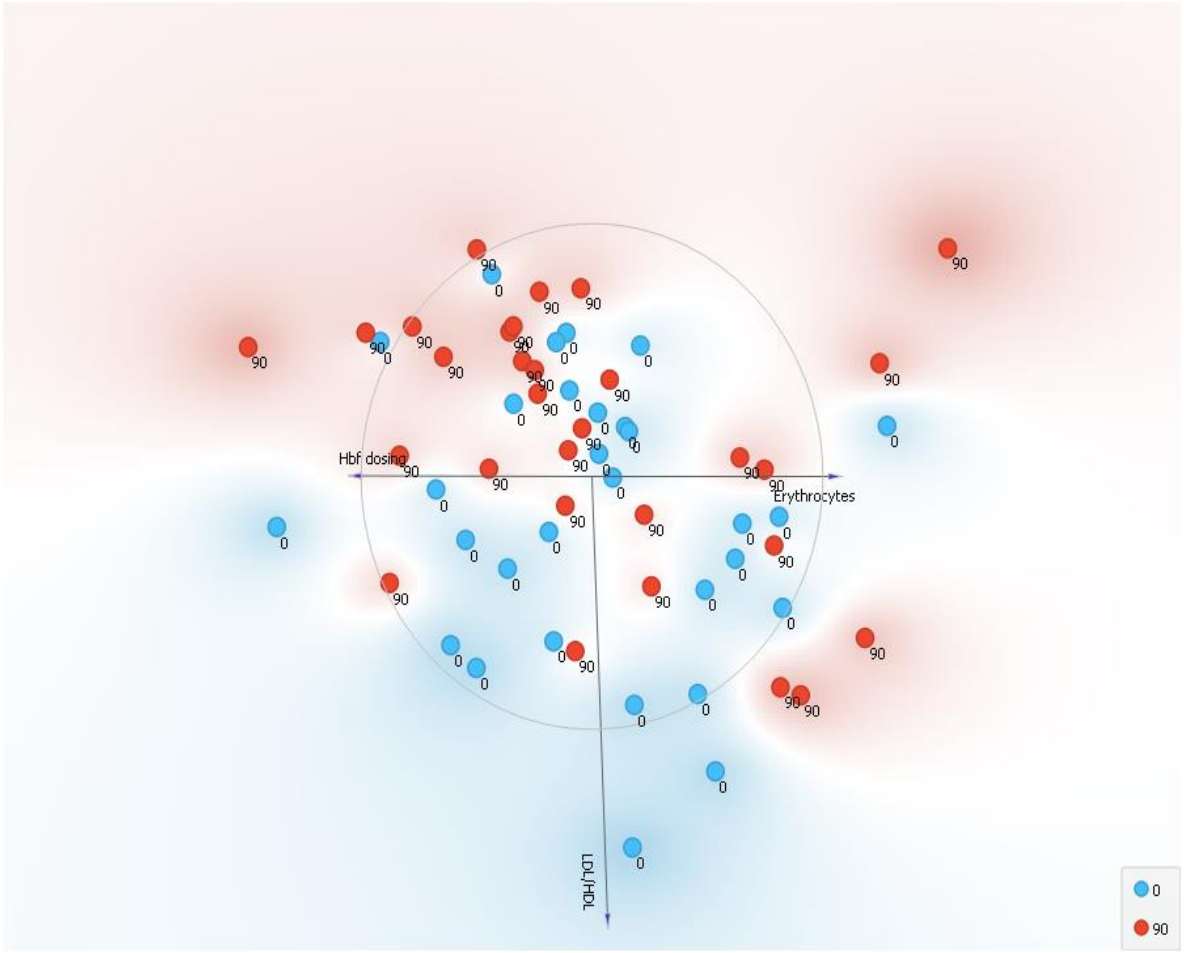


Figure 5.2.2. FreeViz projection of diverse features between T0 (blue) and T90 (red).

The projection cluster in Figure 5.2.2, points of a similar class and repels points of different classes. It is a classification tool. In it, each vector (anchor) represents a feature. The longer it is, the more informative the associated feature. By decreasing or increasing the range of anchors, it was observed that the 3 most relevant features between T0 and T90 were LDL/HDL, erythrocytes and HbF.

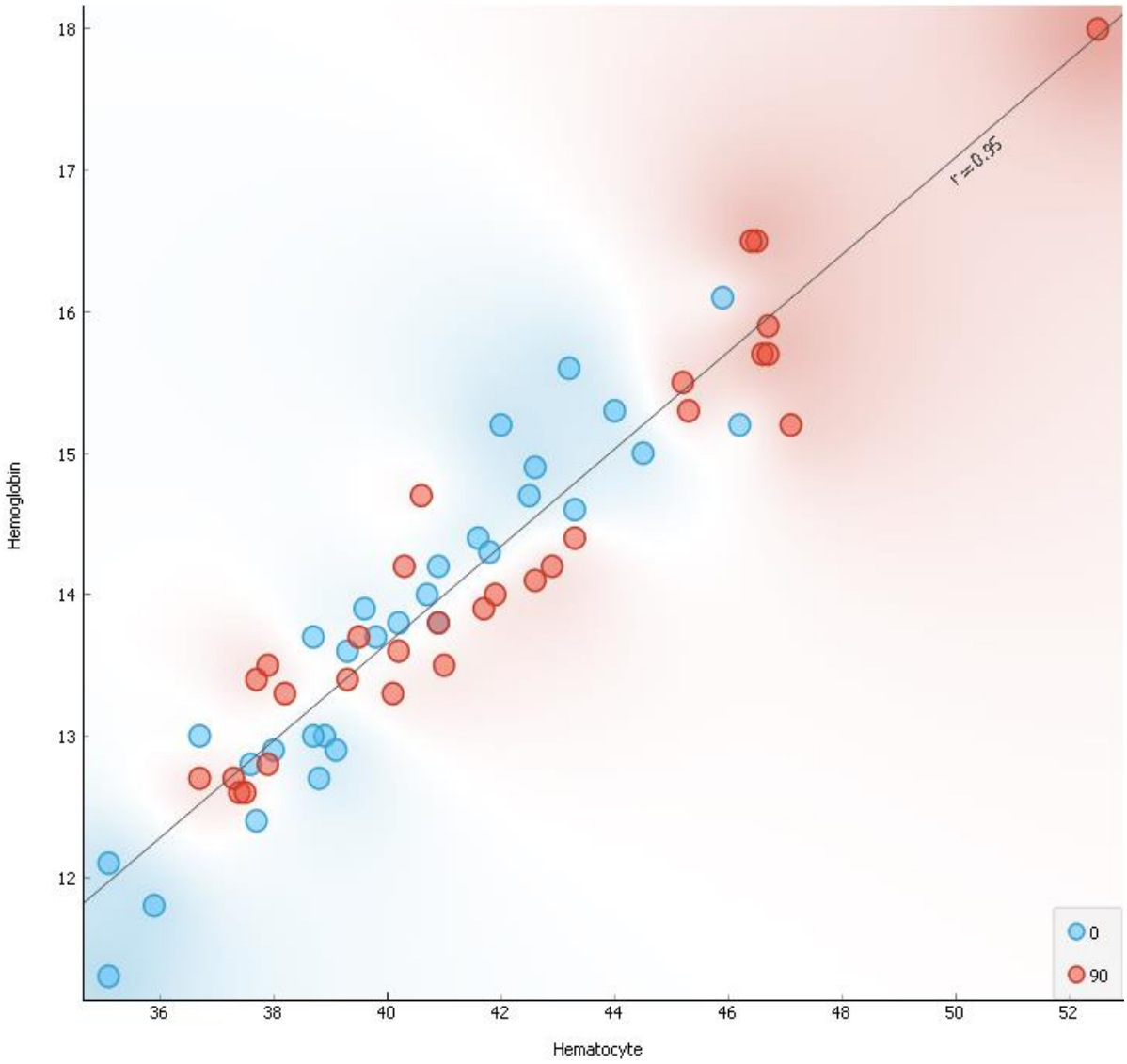


Figure 5.2.3. Spearman's correlations between hematocyte and hemoglobin at T0 (blue) and T90 (red).

On the workflow, there are two correlations to choose from, the Spearman and Pearson. With the present clinical data, similar results were obtained with both types of correlations. As example, Figure 5.2.3 represents the results obtained with the Spearman correlation. The variables presenting higher correlations, (R=0.95) were hematocytes and haemoglobins.

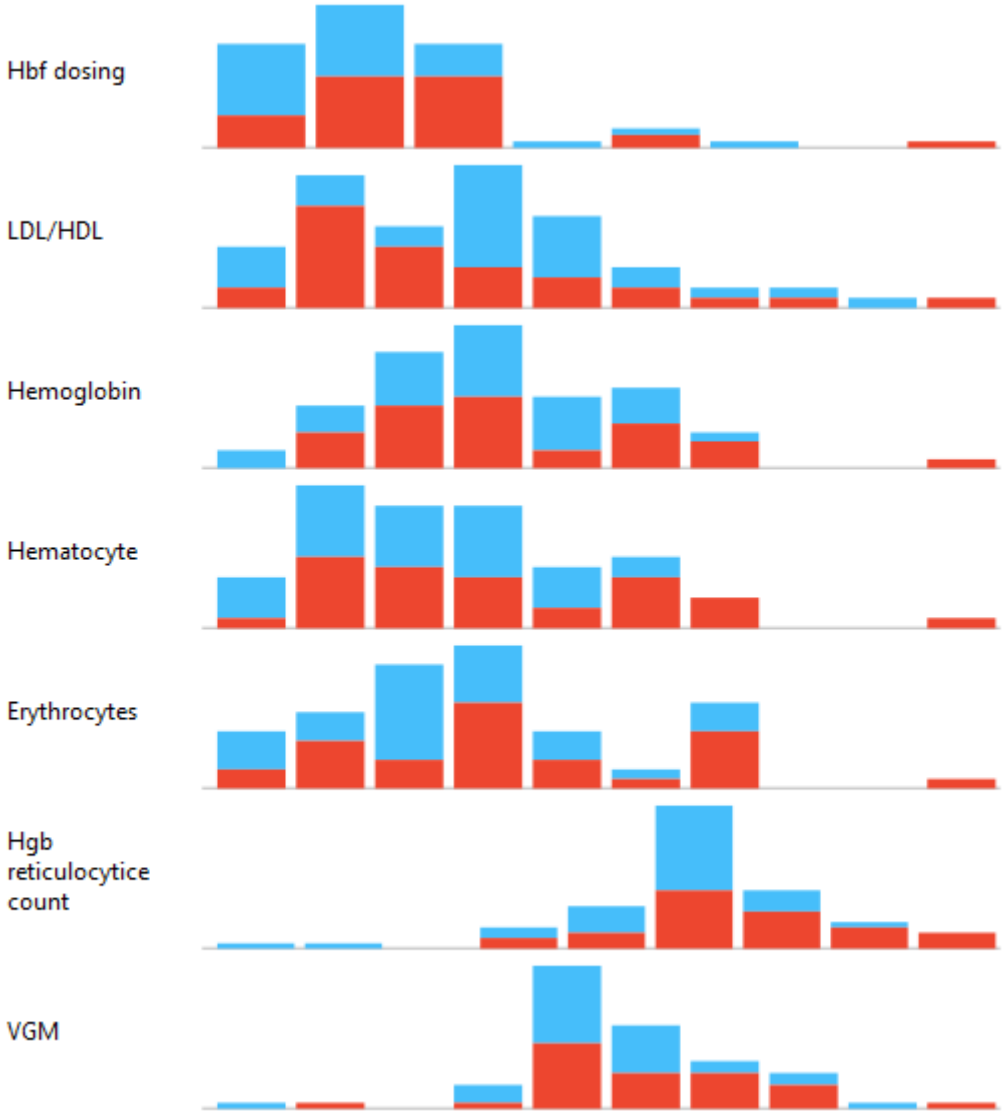


Figure 5.2.4. Feature Statistics.

Feature statistics, as represented in Figure 5.2.4, allows to insight diverse useful information, as the data dispersion [285]. The higher value the better, with HbF being the feature (variable) with the highest dispersion, meaning that the data is well represented in both groups.

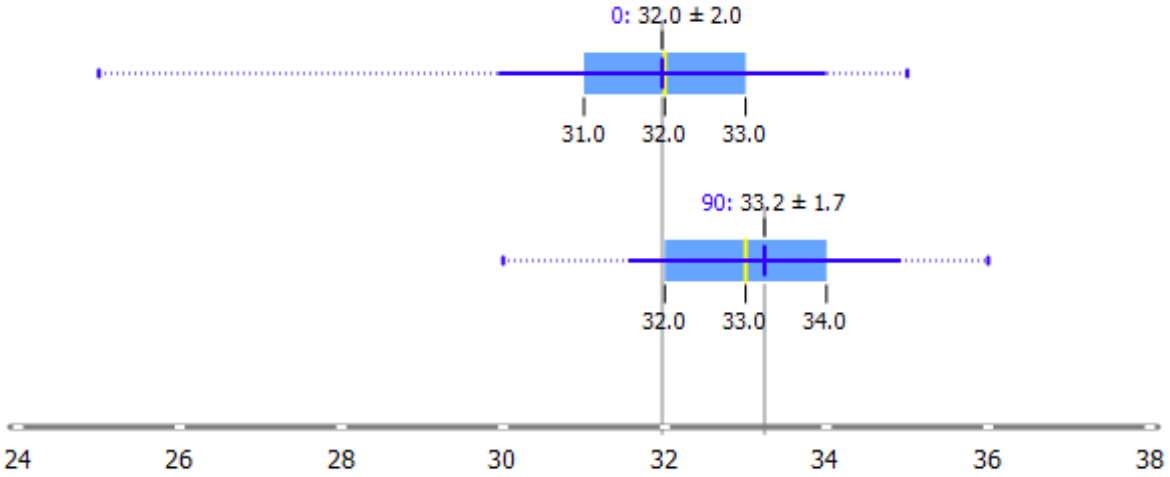


Figure 5.2.5. Box plot for Hgb reticulocyte count at T0 and T90.

The box plot representation (Figure 5.2.5), shows the distribution of attribute values, and is interpreted the following way: the dotted line extends from minimum to maximum value, the vertical blue line is the mean, the horizontal solid blue line indicates one standard deviation from the mean, the yellow line is the median (second quartile) and the blue box spans from the first to the third quartile. Associated to the data, the chi-square statistic was used to order the variables by relevance, showing Hgb reticulocyte count as the one with the more significance difference between T0 and T90. This way, the workflow allowed to conduct inference statistic tests, as *t*-student, resulting in the blood test variables ordered by their *p*-values when comparing T0 with T90 data: hematocyte (p = 0.072), haemoglobin (p = 0.131), VGM (p = 0.144), Hbf dosing (p = 0.163), erythrocytes (p = 0.201), LDL/HDL (p = 0.366).

The tree-viewer representation (Figure 5.2.6), enables the exploratory data analysis and the visualization of classification and regression trees. With the present data it was observed that hematocytes allowed for the best separation, with values below 46.2 belonging to individuals in the T0 group and values above 46.2 belonging to individuals in the T90 group.

The nomogram for the 7 statistically different clinical variables is represented in Figure 5.2.7 and makes use of Naïve Bayes classifiers, a family of probabilistic classifiers based on Bayes' theorem. It offers insight into the structure of the training data and effects of the attributes on class probabilities. It is able to rank the variables by influence or absolute importance (among others), the latter being the ranking method used for this work. For this example, the target class T90 was selected. It is observed, e.g., that for HbF, if a patient presents values between 0.5 and 0.6, the probability of belonging to T90 (i.e., having ingested the EGCG extract during the 90 days) is 100%. If this value was 0.3 or lower than the probability would be of 100% belonging to the T0 group.

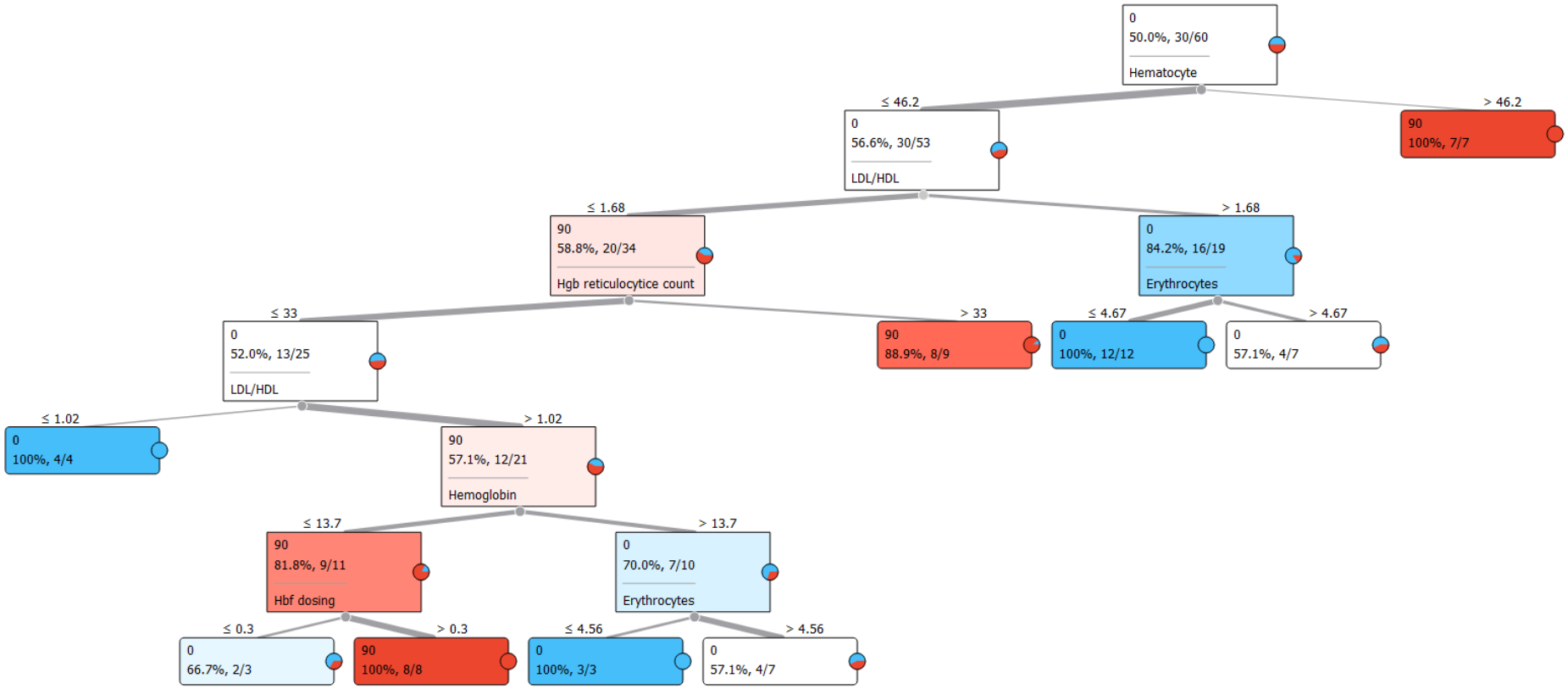


Figure 5.2.6. Tree viewer of 7 conventional clinical variables between T0 and T90.

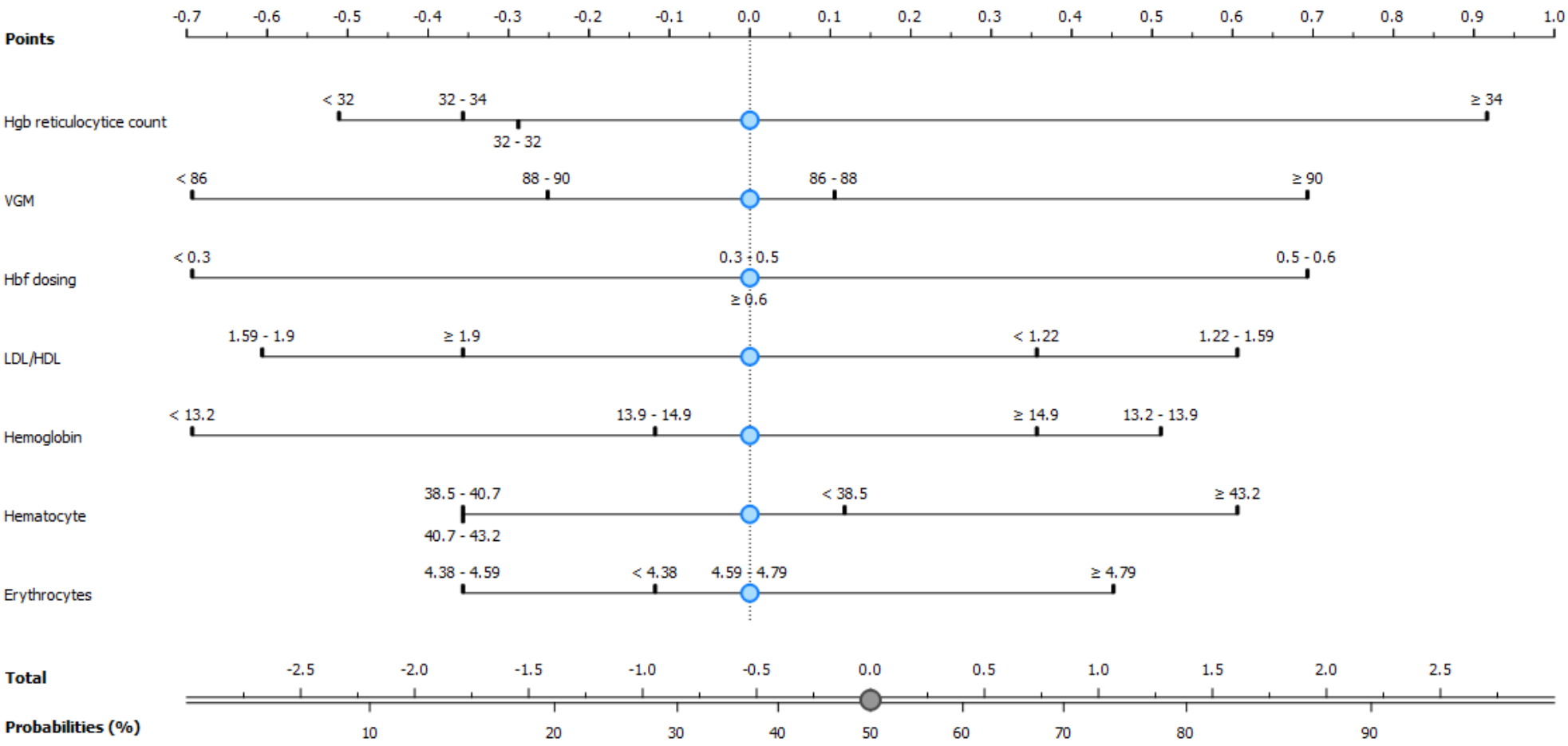


Figure 5.2.7. Nomogram of 7 statistically different clinical variables. Variables scaled by log odds ratios and ranked by absolute importance.

Workflow for spectra data

This workflow (Figure 5.2.8) presents as inputs 4 data matrices: spectra relative to plasma at T0 and T90 and serum spectra at T0 and T90. The dimension of each data matrix is either 90 or 30, in function of using triplicates of the spectra of the 30 participants or the average of the triplicates. For all the cases, input data matrices were spectra pre-processed by atmospheric correction (that compensates atmospheric H_20 and CO_2 at acquisition) by OPUS software. To organize data matrices, Excel was used.

After the initial data matrix is read, the information flows into a test and scores widget and from it to multiple supervised and unsupervised algorithms (e.g. Neural Networks, K- Nearest Neighbors, Naïve Bayes, Support Vector Machines, etc.), with its outputs leading to the *Confusion Matrix*, *Receiver Operating Characteristic* (ROC) curve, *Area Under the Curve* (AUC), *Classification Accuracy* (CA), *F1* (measure of a test's accuracy), *Precision* (also called positive predictive value) and *Recall* (sensitivity). These widgets allow for a comprehensive and visual understanding of the models' classification of data (with labeled data, i.e. for models' calibration) and prediction (with unlabeled data, i.e. for models' validation), plot true positive against false positives in a test and show proportions between the predicted data and real class for each individual tested algorithm.

This is but one of the four ramifications in the workflow, which are as follows:

- **Location 1**, test and scores and referred widgets (Predictions, ROC analysis and Confusion Matrix) are performed on spectral data pre-processed for atmospheric correction;

- **Location 2**, the data has been pre-processed/processed for *Baseline Correction* (rubber band baseline type, positive peak direction, with background subtracted), *Normalization* (Peak from baseline – to Amide I), *Noise reduction* (Gaussian smoothing), *derivative* (Savitzky-Golay filter, 2^{nd} order polynomial, 2^{nd} derivative order, with a 15-points window). Afterwards the pre-processed/processed data is fed into a new data table and the same algorithms are applied and a test and scores results is generated. This data was used for all remaining locations (location 3 and 4);

- **Location 3**, the data passes an additional process called *rank*. Here the spectra are ranked by importance of various statistical tools. ANOVA was chosen. For time constraints, only the top 15 spectra that contribute the most to cluster identification and separation were chosen. This was done as the original spectra has more than 3700 variables, which led to total calculation times between a minimum of 5 hours to hundreds (when calculating spectral correlations). In this location, outliers were not eliminated of the data matrices;

- **Location 4**: In addition to ranking spectra it also dealt with outliers. Outlier detection method was one class SVM with non-linear kernel (RBF) with the Nu parameter set to 50% and a kernel coefficient of 0.01. RBF classifies data as similar or different from the core class (T0 or T90), with the Nu parameter representing an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. Finally, the kernel coefficient specifies how much influence a single data instance has, which is why it is set to the bare minimum, so that each wavenumber has the same weight on the overall spectrum as its neighbor.

Aside from the test and scores results, many other studies are comprised within this workflow, from statistical significance of groups to the minute study of how each individual wavenumber contributes to the overall class separation and which spectral bands correlate with each other the most (absorbance ratios), *visual representations for spectra* (and their averages), *Principal Component Analysis* (PCA) and *Hierarchical Cluster Analysis* (HCA) are all integral parts of our workflow. We present some of these results in subchapter 5.4.

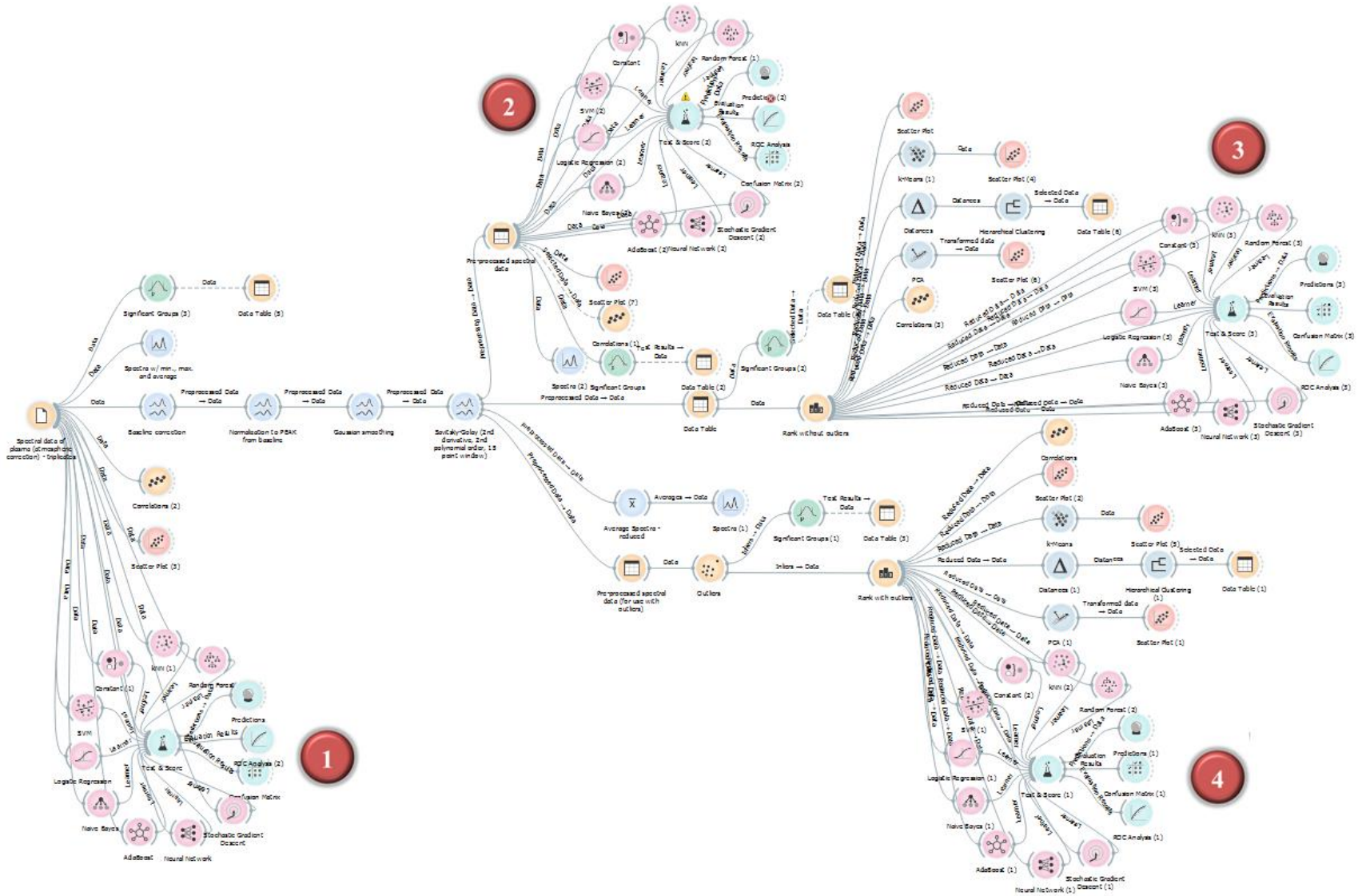


Figure 5.2.8.Main workflows of the study. It is comprised of four separate regions where different pre-processing and processing methods are applied.

5.3. Test & Scores

As to not overburden this chapter, the results of models with spectra from plasma and serum (triplicates and average spectra from replicates), will be presented only in regards of classification accuracy (Table 5.3.2 and Table 5.3.3). As for the parameters used in the test and scores, they were done using Leave One Out (LOO). Calculation times are displayed for the hardware used and referred to in Table 5.3.1. The results revealed a good performance on cluster separation and prediction of unlabeled samples, without showing signs of underfitting or overfitting.

Table 5.3.1. Main hardware and machine settings used.

Operating system	*Windows 8.1 (x64)*
System model	*MSI GT70 2PE*
Processor	*2.8 GHz Intel Core i7-4810MQ*
Cores	*Multi-core (4), Hyper-threaded (8)*
Drive used	*SSD SanDisk SD6SF1M128G*
Graphics card (dedicated)	*NVIDIA GeForce GTX 880M (6GB RAM DDR5)*
Memory Modules	*12 GB RAM DDR5*

The results for the scatter plot informative projections (i.e. calculation of absorbance peak ratios) are, unfortunately, not displayed, as the calculations for all possible combinations, more than 6 million ratios, takes processing times upwards of 50 hours per individual workflow. Since, at this time, exporting the data is not possible, attempts by the user to search specific spectral bands out of the millions of possible combinations, leads to a system crash. In the future, either a substantial hardware upgrade or changes into the software itself, will make this tool relevant for further and continued work.

Table 5.3.2. Test and scores result for all learning methods in the four tested locations in the plasma machine learning workflow. Sampling type: Leave One Out (LOO); Target class: average over classes.

				Tests & Scores – for *classification accuracy*									
Biofluid	**Written designation**	**Location in workflow**	**Calculation time** (*d:h:m:s*)	*Random forest*	*KNN*	*Constant*	*SVM*	*Logistic Regression*	*Naïve Bayes*	*Ada Boost*	*Neural Network*	*Stochastic Gradient Descent*	**Observations**
PLASMA	Raw data	**1**	*(1h:15m)*	0.737	0.872	0.503	0.872	0.978	0.592	0.777	0.955	0.978	In location **1**, the best method was logistic regression. In location **2**, Neural Network had the best overall results. In locations **3** and **4** we show the methods that had perfect results throughout all parameters (AUC, CA, F1, Precision and Recall).
	Pre-processed and processed data	**2**	*(50m)*	0.989	1.000	0.503	0.994	0.497	1.000	0.989	1.000	1.000	
	Pre-processed and processed data, ranked and taking into account outliers only	**3**	*(1m:15s)*	0.986	1.000	0.580	1.000	0.580	1.000	0.993	1.000	1.000	
Subset	Pre-processed and processed data, ranked and not taking into account any outliers	**4**	*(1m:43s)*	1.000	1.000	0.503	1.000	0.503	1.000	0.989	1.000	1.000	
Triplicates													

PLASMA	Raw data	**1**	*(1h:13m)*	0.838	0.872	0.503	0.872	0.978	0.592	0.777	0.972	0.989	In location **1**, the best method was Stochastic Gradient Descent (SGD). In location **2 - 4**, we show highlighted in green the methods that had perfect results throughout all parameters (AUC, CA, F1, Precision and Recall).
	Pre-processed and processed data	**2**	*(47m)*	0.994	1.000	0.503	0.994	0.497	1.000	0.989	1.000	1.000	
	Pre-processed and processed data, ranked and taking into account outliers only	**3**	*(53s)*	0.993	1.000	0.580	1.000	0.580	1.000	0.993	1.000	1.000	
Subset	Pre-processed and processed data, ranked and not taking into account any outliers	**4**	*(1m:15s)*	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	1.000	
Reduced													

Table 5.3.3. Test and scores result for all learning methods in the four tested locations in the serum machine learning workflow. Sampling type: Leave One Out (LOO); Target class: average over classes.

				Tests & Scores – for *classification accuracy*									
Biofluid	**Written designation**	**Location in workflow**	**Calculation time** (*d:h:m:s*)	*Random forest*	*KNN*	*Constant*	*SVM*	*Logistic Regression*	*Naïve Bayes*	*Ada Boost*	*Neural Network*	*Stochastic Gradient Descent*	**Observations**
SERUM	Raw data	**1**	*(1h:07m)*	0.783	0.756	0.000	0.800	0.928	0.611	0.817	0.889	0.956	In location **1**, the best method was stochastic gradient descent (SGD). In location **2 - 4**, we show highlighted in green the methods that had the overall best results and were equal to each other throughout all parameters (AUC, CA, F1, Precision and Recall).
	Pre-processed and processed data	**2**	*(54m)*	0.989	1.000	0.000	1.000	0.000	0.994	0.972	1.000	1.000	
	Pre-processed and processed data, ranked and taking into account outliers only	**3**	*(1m:07s)*	0.994	0.994	0.530	0.988	0.530	0.994	0.982	0.994	0.988	
Subset Triplicates	Pre-processed and processed data, ranked and not taking into account any outliers	**4**	*(1m:36s)*	0.989	0.989	0.000	0.978	0.000	0.989	0.983	0.994	0.994	
SERUM	Raw data	**1**	*(26m)*	0.683	0.667	0.000	0.667	0.867	0.533	0.750	0.917	0.950	In location **1**, the best method was stochastic gradient descent (SGD). In location **2 - 4**, we show highlighted in green the methods that had the overall best results and/or were equal to each other throughout all parameters (AUC, CA, F1, Precision and Recall).
	Pre-processed and processed data	**2**	*(15m:21s)*	1.000	1.000	0.000	0.983	0.000	1.000	0.983	0.983	1.000	
	Pre-processed and processed data, ranked and taking into account outliers only	**3**	*(15s)*	0.935	0.935	0.935	0.935	0.935	0.065	0.806	0.968	0.935	
Subset Reduced	Processed data, ranked and not taking into account any outliers	**4**	*(26s)*	0.983	1.000	0.000	0.983	0.000	1.000	1.000	1.000	1.000	

5.4. Main visual results and chapter conclusion

For simplicity reasons, only outputs not obtained in the previous chapter 4 are highlighted. The workflow starts with the input of pre-processed data with atmospheric correction. It is then displayed the average spectra (solid blue line – T0; solid red line – T90), with all values for T0 and T90 represented as a shadow of the corresponding colour, from its minimum value to its maximum, as represented in Figure 5.4.1. The 2nd derivative from the averaged spectra was also implemented as shown in Figure 5.4.2, as were PCA and HCA, represented in Figure 5.4.3.

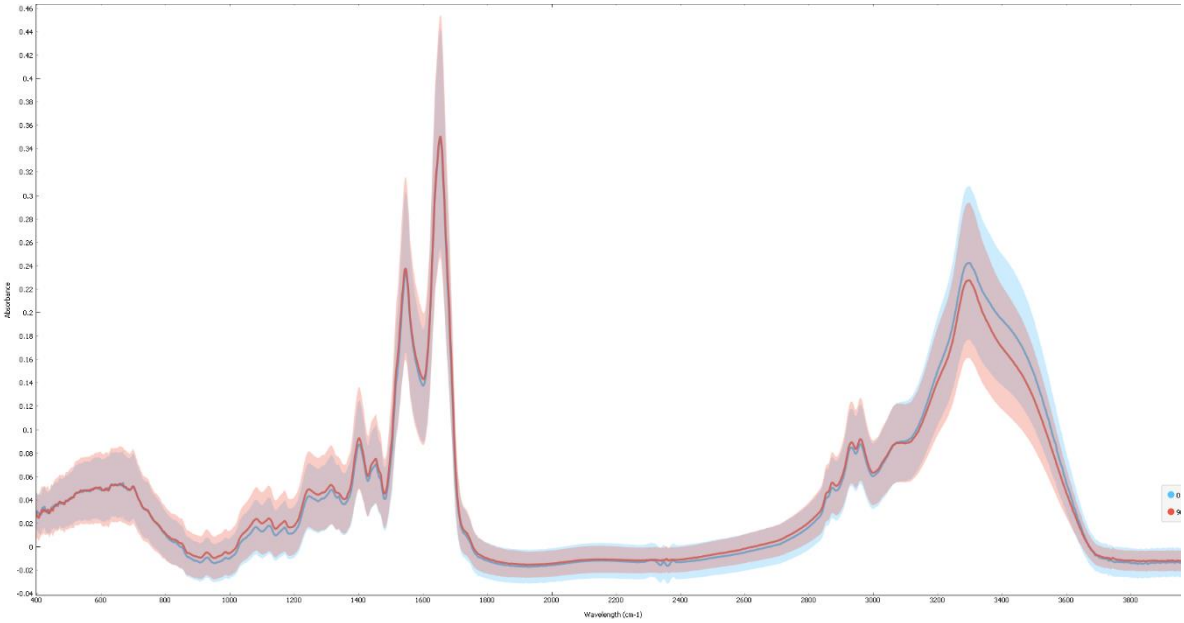


Figure 5.4.1. Spectra of plasma diluted to 1/10 and with atmospheric correction pre-processing. All 180 samples (triplicates) from both T0 group (blue) and T90 group (red) are represented in a colored shadow region, with their corresponding average represented by a solid line of the same color as the represented group.

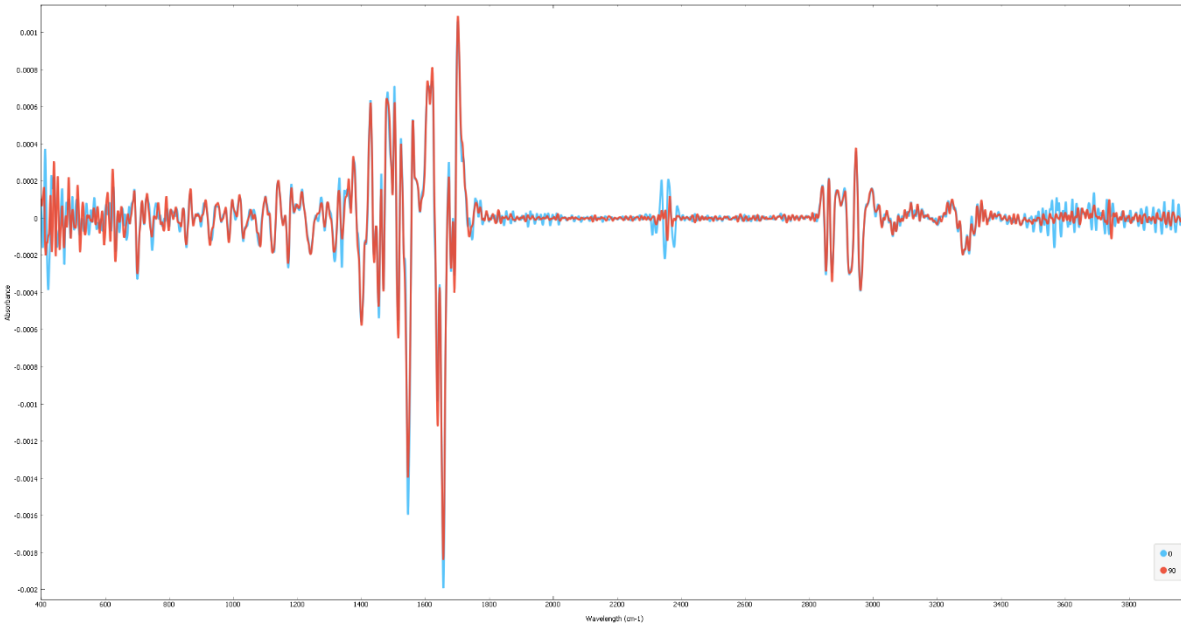


Figure 5.4.2. Second derivative spectra of plasma diluted to 1/10 and with atmospheric and baseline correction and with a normalization to Amide I, Gaussian smoothing and a second derivative, with a Savitzky-Golay filter, a 2nd order polynomial and a 15-points window. The T0 group (blue) and T90 group (red) are represented by a single individual (reduced) spectrum.

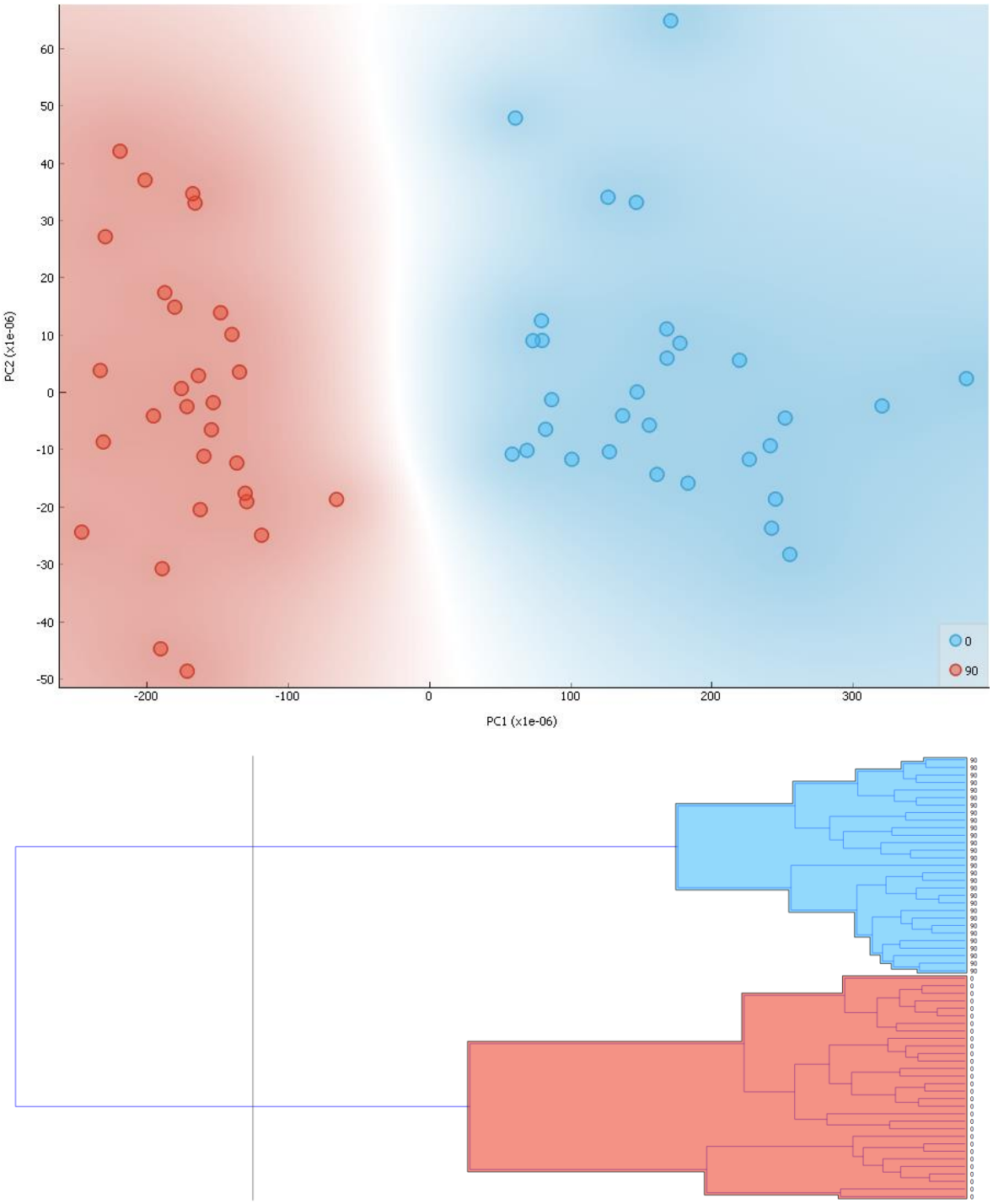


Figure 5.4.3. Above: PCA scatter plot (PC1 vs PC2) for plasma diluted to 1/10 and with atmospheric and baseline correction and with a normalization to Amide I, Gaussian smoothing and a second derivative, with a Savitzky-Golay filter, a 2nd order polynomial and a 15-points window. The data pertains to the reduced data (triplicates averaged). Bellow: HCA of the same pre-processed data. In blue, T0 samples and in red T90 samples, with complete linkage mode.

Concerning the PCA and HCA, the same results were obtained as in chapter 4, albeit with a major boost in result output time and with minimal human input. The results above were for reduced data for plasma and taking into account outliers (**location 3** in the workflow). A perfect group separation between T0 and T90 data on both PCA and HCA were observed.

Having reached the end of this chapter, made incredibly short for data generated and actually studied, I believe that it was represented the capabilities of a workflow that can and it will be continuously improved upon. This workflow enables a faster, more intuitive and visual data representation than the one conducted in chapter 4, which encompassed the use of diverse software and tools along months of data processing and analysis. The Orange based workflow enabled to reduce this time frame to days. The end result being a simple and organic workflow by which, regardless of the spectral data fed into it, allows for automatic pre-processing/processing and output times ranging between minutes to a few days in function of a dataset complexity and study requirements.

There are still a few caveats that need to be addressed in time. As Orange is, in its core, an interactive exploratory data mining tool, it is not yet perfected and lacks a few features (that users can suggest to the open source community), such as: control of the font sizes for exporting images (they are too small when imported into word); lacks full control of outliers, unlike The Unscrambler® X, which undermines somewhat its use into what we wish to consider outliers or not; more advanced widgets like Linear Discriminant Analysis (LDA) which are currently in development (only exist in the more traditional programming line coding). As such, for a complete study we recommend Orange to be used in tandem with specialized software like the ones used throughout this work.

Chapter 6: Conclusions and Future Work

In summary, after daily consumption of 225 mg of EGCG for 90 days on 30 healthy human volunteers, it was possible to observe significant differences at the molecular composition of plasma and serum, between the initial timeframe at the start of the study, before EGCG ingestion (T0) and after the 90 days of consumption (T90). Out of the 35 clinical analysis conducted on the participants blood, the following 7 revealed to be significantly different (at 1%), after the 90 days: erythrocytes, haemoglobin, haematocrit, mean cell volume, reticulocyte haemoglobin content, fetal haemoglobin level and LDL/HDL. From the average MIR spectra (pre-processed with atmospheric correction and a second derivative), of the 30 participants, 48 peaks were detected on plasma and 54 on serum, that were statistically different (at 1% and based on *t*-student), at T90 in relation to T0. Diverse ratios of spectral peaks from plasma and serum were statistically different (at 1% and based on *t*-student), between T90 and T0, that were associated to Amide A, I and II from proteins and CH_2 and CH_3 groups in lipids among others. These results highlight the high impact of EGCG consumption on the plasma and serum molecular profile. These observations are in accordance with other researchers work that observed an impact of EGCG on the general cell metabolism as on metabolic enzyme associated to glycolysis, pentose phosphate pathway and serine biosynthesis [286], mitochondria energetic metabolism [287], lipid metabolism and lipid peroxidation [286], [288], cell division and apoptosis [289].

Traditional and more advanced classification and learning methods associated to FTIRS data and used throughout this work, allowed to observe statistical differences after 90 days of EGCG consumption and identify biomarkers for both biofluids studies (i.e. plasma and serum). The results in the present work are aligned with the increasing applications of machine learning tools and techniques to boon FTIRS biological outputs. While still suffering from some limitations, for example, existence (or lack therefore) of more advanced mathematical operations (e.g., Discriminant Analysis) and hardware capabilities, its future everyday use in laboratories all over the world is one step closer into helping us close the gap on the pipeline between the beginning and end of a study from years to months or even less.

Regardless, there is still much that can be done in this work, from refining the automatic workflows taking into account FTIRS data, program by hand new and more specific tools for spectroscopy to allow the search for biomarkers, determine absorbance ratios for the normal spectrum (and forego the need for extensive data pre-processing), a more extensive study of the clinical variables and effect of EGCG by individual patient, the possibilities are many and exciting.

Chapter 7: Bibliography

[1] J. Trevisan *et al.*, "Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy," *J. Biophotonics*, vol. 7, no. 3–4, pp. 254–265, Apr. 2014.

[2] E. Hoh and V. H. Mair, *The True History of Tea*. Thames & Hudson, 2009.

[3] K. Gascoyne, F. Marchand, J. Desharnais, and H. Américi, *Tea: History, Terroirs, Varieties*, Third. Firefly Books, 2018.

[4] B. Richardson, *The Great Tea Rooms of Britain*, Fifth. Benjamin Press, 2008.

[5] B. Hinsch, *The Ultimate Guide to Chinese Tea*, First. White Lotus Co Ltd, 2008.

[6] Y. Shou-zhong, *The Divine Farmer's Materia Medica*, First. Blue Poppy Pr, 1999.

[7] S. H. Patel, "Camellia sinensis," *J. Agromedicine*, vol. 10, no. 2, pp. 57–64, Oct. 2005.

[8] P. B. Ebrey, *The Cambridge Illustrated History of China*, Second. Cambridge University Press, 2010.

[9] Y. Wang, *Historical Dictionary of Chan Buddhism*. Rowman & Littlefield Publishers, 2017.

[10] C. C. Master and K. Shi, *The Map to Nowhere: Chan Practice Guide to Mind Cultivation*, First. Candlelight Books, 2015.

[11] Bodhidharma and R. Pine, *The Zen Teaching of Bodhidharma*, Bilingual. North Point Press, 1989.

[12] M. Pendergrast, *Uncommon Grounds: The History of Coffee and How It Transformed Our World*, First. Basic Books, 1999.

[13] L. Yu, D. Hitz, and R. Carpenter, *The Classic of Tea*, First. Little, Brown and Company, 1974.

[14] B. Gascoigne, *A Brief History of the Dynasties of China*. Robinson, 2003.

[15] I. Stilwell, *Catherine of Braganza: The courage of a portuguese Infanta who became Queen of England*. Livros Horizonte, 2017.

[16] K. Hubbard, *Serving Victoria: Life in the Royal Household*, Reprint. Harper, 2013.

[17] L. C. Martin, *Tea: The Drink that Changed the World*. Tuttle Publishing, 2007.

[18] H. Su, W. Wu, X. Wan, and J. Ning, "Discriminating geographical origins of green tea based on amino acid, polyphenol, and caffeine content through high-performance liquid chromatography: Taking Lu'an guapian tea as an example," *Food Sci. Nutr.*, vol. 7, no. 6, pp. 2167–2175, 2019.

[19] M. K. Meegahakumbura *et al.*, "Domestication Origin and Breeding History of the Tea Plant (Camellia sinensis) in China and India Based on Nuclear Microsatellites and cpDNA Sequence Data," *Front. Plant Sci.*, vol. 8, no. January, pp. 1–12, 2018.

[20] C. S. Yang, G. Chen, and Q. Wu, "Recent Scientific Studies of a Traditional Chinese Medicine, Tea, on Prevention of Chronic Diseases," *J. Tradit. Complement. Med.*, vol. 4, no. 1, pp. 17–23, 2014.

[21] A. C. Graham, *Studies in Chinese Philosophy and Philosophical Literature: Logic and Reality*. SUNY Press, 1986.

[22] J. H. Weisburger, "Tea and health: A historical perspective," *Cancer Lett.*, vol. 114, no. 1–2, pp. 315–317, 1997.

[23] "PubMed." [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed?term=green tea%5BTitle%2FAbstract%5D. [Accessed: 08-Aug-2019].

[24] Euromonitor, "Tea Global Corporate Strategy: Diversity and Tea Experience," 2015.

[25] C. Ruxton, F. Phillips, and T. Bond, "Is tea a healthy source of hydration?," *Nutr. Bull.*, vol. 40, no. 3, pp. 166–176, 2015.

[26] F. and A. O. of the U. Nations, "World tea production and trade - current and future development," 2016. [Online]. Available: http://www.fao.org/3/a-i4480e.pdf.

[27] D. Grigg, "The worlds of tea and coffee: Patterns of consumption," *GeoJournal*, vol. 57, no. 4, pp. 283–294, 2002.

[28] N. Hall, *The Tea Industry*, First. Woodhead Publishing, 2000.

[29] E. Rappaport, *A Thirst for Empire: How Tea Shaped the Modern World*. Princeton University Press, 2017.

[30] A. Nehlig, J. L. Daval, and G. Debry, “Caffeine and the central nervous system: mechanisms of action.,” *Brain Res. Rev.*, vol. 17, no. 2, pp. 139–170, 1992.
[31] T. I. A. for R. on Cancer, *Coffee, Tea, Mate, Methylxanthines and Methylglyoxal*, First. World Health Organization, 1991.
[32] N. C. for B. Information, “L-Theanine, CID=439378,” *PubChem Database*. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/L-Theanine. [Accessed: 09-Aug-2019].
[33] D. C. Chu, T. Okubo, Y. Nagato, and H. Yokogoshi, “L-theanine - A unique amino acid of green tea and its relaxation effect in humans,” *Trends Food Sci. Technol.*, vol. 10, no. 6–7, pp. 199–204, 1999.
[34] V. Dramard *et al.*, “Effect of l-theanine tablets in reducing stress-related emotional signs in cats: An open-label field study,” *Ir. Vet. J.*, vol. 71, no. 1, pp. 7–9, 2018.
[35] A. C. Nobre, A. Rao, and G. N. Owen, “L-theanine, a natural constituent in tea, and its effect on mental state.: Discovery Service for Endeavour College of Natural Health Library,” *Asia Pac. J. Clin. Nutr.*, vol. 17, no. S1, pp. 167–168, 2008.
[36] R. Huang, A. J. O’Donnell, J. J. Barboline, and T. J. Barkman, “Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 38, pp. 10613–10618, 2016.
[37] B. Stavric, R. Klassen, B. Watkinson, K. Karpinski, R. Stapley, and P. Fried, “Variability in caffeine consumption from coffee and tea: Possible significance for epidemiological studies,” *Food Chem. Toxicol.*, vol. 26, no. 2, pp. 111–118, 1988.
[38] S. Khokhar and S. G. M. Magnusdottir, “Total phenol, catechin, and caffeine contents of teas commonly consumed in the United Kingdom,” *J. Agric. Food Chem.*, vol. 50, no. 3, pp. 565–570, 2002.
[39] J. M. Chin, M. L. Michelle, B. A. Goldbergeer, S.-C. Angela, and E. J. Cone, “Caffeine content in brewed teas,” *J. Anal. Toxicol.*, vol. 32, no. October, pp. 702–704, 2008.
[40] R. M. Gilbert, J. A. Marshman, M. Schwieder, and R. Berg, “Caffeine content of beverages as consumed,” *Can. Med. Assoc. J.*, vol. 114, no. 3, pp. 205–208, 1976.
[41] J. L. Temple, C. Bernard, S. E. Lipshultz, J. D. Czachor, J. A. Westphal, and M. A. Mestre, “The Safety of Ingested Caffeine: A Comprehensive Review,” *Front. Psychiatry*, vol. 8, no. May, pp. 1–19, 2017.
[42] I. and C. C. in L. M. of N. Sharma, HemrajEstimation of Caffeine Content on Various Brands of Tea of Nepal, H. P. Sapkota, S. Khan, and S. B. Bogati, “Estimation of Caffeine Content on Various Brands of Tea of Nepal , India and China Consumed in Local Market of Nepal,” *Inven. Rapid Pharm Anal. Qual. Assur.*, no. June, 2019.
[43] A. Gramza, “Caffeine in Tea Camellia Sinensis – Content , Absorption ,” vol. 18, no. 2, 2014.
[44] A. Akula and C. Akula, “Somatic Embryogenesis in Tea (Camellia sinensis (L.) O. Kuntze),” in *Forestry Sciences (vol.59)*, Springer, Dordrecht, 1999, pp. 239–257.
[45] S. Taylor, “Encyclopedia of Food Sciences and Nutrition,” *Encycl. Food Sci. Nutr.*, pp. 5737–5743, 2003.
[46] R. J. Heiss, *The Story of Tea: A Cultural History and Drinking Guide*, First. Ten Speed Press, 2007.
[47] W. Battle, *The World Tea Encyclopaedia: The World of Tea Explored and Explained from Bush to Brew*, UK. Matador, 2017.
[48] K.-H. Hong and A. of T. Brother, *The Korean Way of Tea: An Introductory Guide*, First. Seoul Selection USA, Inc., 2011.
[49] M. Reto, M. E. Figueira, H. M. Filipe, and C. M. M. Almeida, “Chemical composition of green tea (Camellia sinensis) infusions commercialized in Portugal,” *Plant Foods Hum. Nutr.*, vol. 62, no. 4, pp. 139–144, 2007.
[50] P. V., A. M., D. S., H. G., and H. S.K., “A review on: Green tea: A miraculous drink,” *Int. J. Pharm. Sci. Rev. Res.*, vol. 51, no. 2, pp. 26–34, 2018.
[51] C. S.M., T. P.T., K. R., and N. I., “Beneficial effects of green tea: A literature review,” *Chin. Med.*, vol. 5, pp. 1–9, 2010.
[52] C. C., A. R., and G. R., “Beneficial effects of green tea - A review,” *J. Am. Coll. Nutr.*, vol. 25, no. 2, pp. 79–99, 2006.
[53] Y. SUZUKI, N. MIYOSHI, and M. ISEMURA, “Health-promoting effects of green tea,” *Proc.*

Japan Acad. Ser. B, vol. 88, no. 3, pp. 88–101, 2012.
[54] M. H. Pan, Y. S. Chiou, Y. J. Wang, C. T. Ho, and J. K. Lin, "Multistage carcinogenesis process as molecular targets in cancer chemoprevention by epicatechin-3-gallate," *Food Funct.*, vol. 2, no. 2, pp. 101–110, 2011.
[55] A. Manuscript, "Nihms585940.Pdf," vol. 19, no. 34, pp. 6141–6147, 2014.
[56] Y. Kuroda and Y. Hara, "Antimutagenic and anticarcinogenic activity of tea polyphenols," *Mutat. Res. - Rev. Mutat. Res.*, vol. 436, no. 1, pp. 69–97, 1999.
[57] M. W. L. Koo and C. H. Cho, "Pharmacological effects of green tea on the gastrointestinal system," *Eur. J. Pharmacol.*, vol. 500, no. 1-3 SPEC. ISS., pp. 177–185, 2004.
[58] P. Sciences, "No. 9] Proc. Japan Acad., 78, Ser. B (2002) 263," no. 9, pp. 263–270, 2002.
[59] Z. Chen and Z. Lin, "Tea and human health: biomedical functions of tea active components and current issues," *J. Zhejiang Univ. B*, vol. 16, no. 2, pp. 87–102, Feb. 2015.
[60] W. C. Reygaert, "Green tea catechins: Their use in treating and preventing infectious diseases," *Biomed Res. Int.*, vol. 2018, 2018.
[61] C. S. Yang and H. Wang, "Mechanistic issues concerning cancer prevention by tea catechins," *Mol. Nutr. Food Res.*, vol. 55, no. 6, pp. 819–831, 2011.
[62] S. Wolfram, "Effects of green tea and egcg on cardiovascular and metabolic health," *J. Am. Coll. Nutr.*, vol. 26, no. 4, pp. 373S-388S, 2007.
[63] J. A. Vinson and Y. A. Dabbagh, "Effect of green and black tea supplementation on lipids, lipid oxidation and fibrinogen in the hamster: Mechanisms for the epidemiological benefits of tea drinking," *FEBS Lett.*, vol. 433, no. 1–2, pp. 44–46, 1998.
[64] H. TACHIBANA, "Green tea polyphenol sensing," *Proc. Japan Acad. Ser. B*, vol. 87, no. 3, pp. 66–80, 2011.
[65] C.-L. Sun, "Urinary tea polyphenols in relation to gastric and esophageal cancers: a prospective study of men in Shanghai, China," *Carcinogenesis*, vol. 23, no. 9, pp. 1497–1503, 2002.
[66] M. Suganuma, A. Saha, and H. Fujiki, "New cancer treatment strategy using combination of green tea catechins and anticancer drugs," *Cancer Sci.*, vol. 102, no. 2, pp. 317–323, 2011.
[67] M. Shimizu, S. Adachi, M. Masuda, O. Kozawa, and H. Moriwaki, "Cancer chemoprevention with green tea catechins by targeting receptor tyrosine kinases," *Mol. Nutr. Food Res.*, vol. 55, no. 6, pp. 832–843, 2011.
[68] M. Noguchi-Shinohara *et al.*, "Consumption of green tea, but not black tea or coffee, is associated with reduced risk of cognitive decline," *PLoS One*, vol. 9, no. 5, 2014.
[69] K. Niu *et al.*, "Green tea consumption is associated with depressive symptoms in the elderly," *Am. J. Clin. Nutr.*, vol. 90, no. 6, pp. 1615–1622, Dec. 2009.
[70] T. J.R. and W. V.M., "Probable antagonism of warfarin by green tea," *Ann. Pharmacother.*, vol. 33, no. 4, pp. 426–428, 1999.
[71] R. Soltani, A. Haghighat, M. Fanaei, and G. Asghari, "Evaluation of the effect of green tea extract on the prevention of gingival bleeding after posterior mandibular teeth extraction: A randomized controlled trial," *Evidence-based Complement. Altern. Med.*, vol. 2014, 2014.
[72] T. O. Cheng, "Green tea may inhibit warfarin," *Int. J. Cardiol.*, vol. 115, no. 2, p. 236, 2007.
[73] M. I. Prasanth, B. S. Sivamaruthi, C. Chaiyasut, and T. Tencomnao, "A review of the role of green tea (camellia sinensis) in antiphotoaging, stress resistance, neuroprotection, and autophagy," *Nutrients*, vol. 11, no. 2, 2019.
[74] K. Han, E. Hwang, and J. B. Park, "Excessive consumption of green tea as a risk factor for periodontal disease among korean adults," *Nutrients*, vol. 8, no. 7, pp. 1–10, 2016.
[75] A. M. Dostal *et al.*, "The safety of green tea extract supplementation in postmenopausal women at risk for breast cancer: results of the Minnesota Green Tea Trial," *Food Chem. Toxicol.*, vol. 83, no. 3, pp. 26–35, Sep. 2015.
[76] T. Isomura *et al.*, "Liver-related safety assessment of green tea extracts in humans: A systematic review of randomized controlled trials," *Eur. J. Clin. Nutr.*, vol. 70, no. 11, pp. 1221–1229, 2016.
[77] M. N. Mead, "Temperance in green tea.," *Environ. Health Perspect.*, vol. 115, no. 9, pp. 444–447, 2007.
[78] S. J. K. Chong, K. A. Howard, and C. Knox, "Hypokalaemia and drinking green tea: A literature review and report of 2 cases," *BMJ Case Rep.*, vol. 2016, pp. 4–5, 2016.
[79] M. Younes *et al.*, "Scientific opinion on the safety of green tea catechins," *EFSA J.*, vol. 16,

no. 4, 2018.
[80] B. Halliwell, “Are polyphenols antioxidants or pro-oxidants? What do we learn from cell culture and in vivo studies?,” *Arch. Biochem. Biophys.*, vol. 476, no. 2, pp. 107–112, 2008.
[81] U. S. N. L. of M. NIH, “Green tea,” *Toxicoloy Data Network*. [Online]. Available: https://toxnet.nlm.nih.gov/cgi-bin/sis/search2/r?dbs+hsdb:@term+@na+Green+tea. [Accessed: 11-Aug-2019].
[82] U. S. N. L. of M. NIH, “Green Tea studies.” [Online]. Available: https://clinicaltrials.gov/ct2/results?term=green+tea. [Accessed: 17-Aug-2019].
[83] M. Grzesik, K. Naparło, G. Bartosz, and I. Sadowska-Bartosz, “Antioxidant properties of catechins: Comparison with other antioxidants,” *Food Chem.*, vol. 241, pp. 480–492, 2018.
[84] A. Zinellu *et al.*, “Human serum albumin increases the stability of green tea catechins in aqueous physiological conditions,” *PLoS One*, vol. 10, no. 7, pp. 1–12, 2015.
[85] P. L. Toutain and A. Bousquet-Mélou, “Bioavailability and its assessment,” *J. Vet. Pharmacol. Ther.*, vol. 27, no. 6, pp. 455–466, 2004.
[86] I. S. Rombauer, *Joy of Cooking*, Anniversar. Scribner, 2006.
[87] M. Serafini, A. Ghiselli, and A. Ferro-Luzzi, “In vivo antioxidant effect of green and black tea in man.,” *Eur. J. Clin. Nutr.*, vol. 50, no. 1, pp. 28–32, Jan. 1996.
[88] S. C. Langley-Evans, “Consumption of black tea elicits an increase in plasma antioxidant potential in humans,” *Int. J. Food Sci. Nutr.*, vol. 51, no. 5, pp. 309–315, 2000.
[89] J. A. M. Kyle, P. C. Morrice, G. McNeill, and G. G. Duthie, “Effects of infusion time and addition of milk on content and absorption of polyphenols from black tea,” *J. Agric. Food Chem.*, vol. 55, no. 12, pp. 4889–4894, 2007.
[90] N. H. Lee *et al.*, “Bioavailability and antioxidant activity of tea flavanols after consumption of green tea, black tea, or a green tea extract supplement,” *Am. J. Clin. Nutr.*, vol. 80, no. 6, pp. 1558–1564, 2018.
[91] K. H. Van Het Hof, G. A. A. Kivits, J. A. Weststrate, and L. B. M. Tijburg, “Bioavailability of catechins from tea: The effect of milk,” *Eur. J. Clin. Nutr.*, vol. 52, no. 5, pp. 356–359, 1998.
[92] C. Manach, G. Williamson, C. Morand, A. Scalbert, and C. Rémésy, “Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies,” *Am. J. Clin. Nutr.*, vol. 81, no. 1, pp. 230S-242S, Jan. 2005.
[93] A. Kale *et al.*, “Studies on the effects of oral administration of nutrient mixture, quercetin and red onions on the bioavailability of epigallocatechin gallate from green tea extract,” *Phyther. Res.*, vol. 24, no. S1, pp. S48–S55, Jan. 2010.
[94] B. M. Ryan, T. J. Dougherty, D. Beaulieu, J. Chuang, B. A. Dougherty, and J. F. Barrett, “Efflux in bacteria: what do we really know about it?,” *Expert Opin. Investig. Drugs*, vol. 10, no. 8, pp. 1409–1422, 2005.
[95] J. Zhang, S. Nie, and S. Wang, “Nanoencapsulation enhances epigallocatechin-3-gallate stability and its antiatherogenic bioactivities in macrophages,” *J. Agric. Food Chem.*, vol. 61, no. 38, pp. 9200–9209, 2013.
[96] J. Zhang, S. Nie, R. Martinez-Zaguilan, S. R. Sennoune, and S. Wang, “Formulation, characteristics and antiatherogenic bioactivities of CD36-targeted epigallocatechin gallate (EGCG)-loaded nanoparticles,” *J. Nutr. Biochem.*, vol. 30, pp. 14–23, 2016.
[97] D. Mereles and W. Hunstein, “Epigallocatechin-3-gallate (EGCG) for clinical trials: More Pitfalls than Promises?,” *Int. J. Mol. Sci.*, vol. 12, no. 9, pp. 5592–5603, 2011.
[98] S. K. Ramaiah and A. Banerjee, *Liver Toxicity of Chemical Warfare Agents*. Elsevier Inc., 2015.
[99] S. M. Henning, J. J. Choo, and D. Heber, “Nongallated Compared with Gallated Flavan-3-ols in Green and Black Tea Are More Bioavailable,” *J. Nutr.*, vol. 138, no. 8, pp. 1529S-1534S, 2008.
[100] W. Yong Feng, “Metabolism of Green Tea Catechins: An Overview,” *Curr. Drug Metab.*, vol. 7, no. 7, pp. 755–809, 2006.
[101] A. Diniz, L. Escuder-Gilabert, N. P. Lopes, R. M. Villanueva-Camañas, S. Sagrado, and M. J. Medina-Hernández, “Characterization of interactions between polyphenolic compounds and human serum proteins by capillary electrophoresis,” *Anal. Bioanal. Chem.*, vol. 391, no. 2, pp. 625–632, 2008.
[102] S. M. Henning *et al.*, “Bioavailability and antioxidant effect of epigallocatechin gallate

administered in purified form versus as green tea extract in healthy individuals," *J. Nutr. Biochem.*, vol. 16, no. 10, pp. 610–616, 2005.
[103] D. J. Boocock *et al.*, "Phase I pharmacokinetic study of tea polyphenols following single-dose administration of epigallocatechin gallate and Polyphenon E," *Cancer Epidemiol. Biomarkers Prev.*, vol. 16, no. 6, pp. 1246–1252, 2007.
[104] T. Unno, K. Kondo, H. Itakura, and T. Takeo, "Analysis of (–)-Epigallocatechin Gallate in Human Serum Obtained after Ingesting Green Tea," *Biosci. Biotechnol. Biochem.*, vol. 60, no. 12, pp. 2066–2068, 1996.
[105] R. D.E. and K. C.D., "Biliary excretion of drugs in man," *Clin. Pharmacokinet.*, vol. 4, no. 5, pp. 368–379, 1979.
[106] C. S. Yang, L. Chen, M.-J. Lee, D. Balentine, M. C. Kuo, and S. P. Schantz, "Blood and urine levels of tea catechins after ingestion of different amounts of green tea by human volunteers," *Cancer Epidemiol Biomarkers Prev.*, vol. 7, pp. 351–354, 1998.
[107] B. Yang, K. Arai, and F. Kusu, "Determination of catechins in human urine subsequent to tea ingestion by high-performance liquid chromatography with electrochemical detection," *Anal. Biochem.*, vol. 283, no. 1, pp. 77–82, 2000.
[108] K. Strimbu and J. A. Tavel, "What are biomarkers?," *Curr. Opin. HIV AIDS*, vol. 5, no. 6, pp. 463–466, 2010.
[109] FDA-NIH, "BEST (Biomarkers, EndpointS, and other Tools) Resource," *Natl. Institutes Heal.*, 2018.
[110] R. Mayeux, "Biomarkers: Potential uses and limitations," *NeuroRX*, vol. 1, no. 2, pp. 182–188, Apr. 2004.
[111] R. M. Califf, "Biomarker definitions and their applications," *Exp. Biol. Med.*, vol. 243, no. 3, pp. 213–221, 2018.
[112] J. Deelen *et al.*, "A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals," *Nat. Commun.*, vol. 10, no. 1, p. 3346, Dec. 2019.
[113] M.-C. Yu *et al.*, "Label Free Detection of Sensitive Mid-Infrared Biomarkers of Glomerulonephritis in Urine Using Fourier Transform Infrared Spectroscopy," *Sci. Rep.*, vol. 7, no. 1, p. 4601, Dec. 2017.
[114] S. Kar, D. R. Katti, and K. S. Katti, "Fourier transform infrared spectroscopy based spectral biomarkers of metastasized breast cancer progression," *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, vol. 208, pp. 85–96, Feb. 2019.
[115] D. I. Ellis, G. G. Harrigan, and R. Goodacre, "Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy," in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Boston, MA: Springer US, 2003, pp. 111–124.
[116] L. Calamari, A. Ferrari, A. Minuti, and E. Trevisi, "Assessment of the main plasma parameters included in a metabolic profile of dairy cow based on Fourier Transform mid-infrared spectroscopy: preliminary results," *BMC Vet. Res.*, vol. 12, no. 1, p. 4, Dec. 2016.
[117] F. Matin *et al.*, "A Plasma Biomarker Panel of Four MicroRNAs for the Diagnosis of Prostate Cancer," *Sci. Rep.*, vol. 8, no. 1, p. 6653, Dec. 2018.
[118] M. Le Corvec *et al.*, "Mid-infrared spectroscopy of serum, a promising non-invasive method to assess prognosis in patients with ascites and cirrhosis," *PLoS One*, vol. 12, no. 10, p. e0185997, Oct. 2017.
[119] D. M. Rathod, K. R. Patel, H. N. Mistri, A. G. Jangid, P. S. Shrivastav, and M. Sanyal, "Simultaneous analysis of allopurinol and oxypurinol using a validated liquid chromatography–tandem mass spectrometry method in human plasma," *J. Pharm. Anal.*, vol. 7, no. 1, pp. 56–62, 2017.
[120] M. M. Tibben *et al.*, "Liquid chromatography-tandem mass spectrometric assay for the quantification of galunisertib in human plasma and the application in a pre-clinical study," *J. Pharm. Biomed. Anal.*, vol. 173, pp. 169–175, 2019.
[121] D. Spaggiari *et al.*, "Development and validation of a multiplex UHPLC-MS/MS method for the determination of the investigational antibiotic against multi-resistant tuberculosis macozinone (PBTZ169) and five active metabolites in human plasma," *PLoS One*, vol. 14, no. 5, p. e0217139, May 2019.
[122] J. Lan *et al.*, "Systematic Evaluation of the Use of Human Plasma and Serum for Mass-Spectrometry-Based Shotgun Proteomics," *J. Proteome Res.*, vol. 17, no. 4, pp. 1426–1435,

Apr. 2018.
[123] N. Liu, E. Tengstrand, K. O. Boernsen, S. Bek, and F. Hsieh, “Validation of a multiplexed LC–MS/MS clinical assay to quantify insulin-like growth factor-binding proteins in human serum and its application in a clinical study,” *Toxicol. Appl. Pharmacol.*, vol. 371, pp. 74–83, 2019.
[124] A. J. Boggess, G. M. M. Rahman, M. Pamukcu, S. Faber, and H. M. S. Kingston, “An accurate and transferable protocol for reproducible quantification of organic pollutants in human serum using direct isotope dilution mass spectrometry,” *Analyst*, vol. 139, no. 23, pp. 6223–6231, 2014.
[125] J. Moon, J. H. Ko, C. H. Yoon, M. K. Kim, and J. Y. Oh, “Effects of 20% Human Serum on Corneal Epithelial Toxicity Induced by Benzalkonium Chloride: In Vitro and Clinical Studies,” *Cornea*, vol. 37, no. 5, pp. 617–623, 2018.
[126] A. S. Jaffe *et al.*, “It’s time for a change to a troponin standard,” *Circulation*, vol. 102, no. 11, pp. 1216–1220, 2000.
[127] D. B. Sacks, D. E. Bruns, D. E. Goldstein, N. K. Maclaren, J. M. McDonald, and M. Parrott, “Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus.,” *Clin. Chem.*, vol. 48, no. 3, pp. 436–72, Mar. 2002.
[128] C. Petibois and G. Déléris, “2D-FT-IR spectrometry: A new tool for the analysis of stress-induced plasma content changes,” *Vib. Spectrosc.*, vol. 32, no. 1 SPEC., pp. 117–128, 2003.
[129] Z. Yu *et al.*, “Differences between human plasma and serum metabolite profiles,” *PLoS One*, vol. 6, no. 7, pp. 1–6, 2011.
[130] “Difference between Blood Plasma and Serum.” [Online]. Available: https://www.easybiologyclass.com/difference-between-blood-plasma-and-serum/. [Accessed: 24-Aug-2019].
[131] P. Pokhrel, “Differences between Serum and Plasma,” *Microbiology notes*. [Online]. Available: http://microbiologynotes.com/differences-between-serum-and-plasma/. [Accessed: 24-Aug-2019].
[132] S. Aryal, “Difference between Serum and Plasma,” 2019. [Online]. Available: https://microbiologyinfo.com/difference-between-serum-and-plasma/. [Accessed: 24-Aug-2019].
[133] J. H. Austin, “Blood: Physiology of Formed Elements and Plasma; Blood Clotting,” *Annu. Rev. Physiol.*, vol. 1, no. 1, pp. 297–316, Mar. 1939.
[134] H. A. Krebs, “Chemical Composition of Blood Plasma and Serum,” *Annu. Rev. Biochem.*, vol. 19, no. 1, pp. 409–430, 1950.
[135] C. Oddoze, E. Lombard, and H. Portugal, “Stability study of 81 analytes in human whole blood, in serum and in plasma,” *Clin. Biochem.*, vol. 45, no. 6, pp. 464–469, 2012.
[136] N. Psychogios *et al.*, “The human serum metabolome,” *PLoS One*, vol. 6, no. 2, 2011.
[137] W. H. O. D. I. and L. Technology, “Use of anticoagulants in diagnostic laboratory investigations,” 2002.
[138] A. DiBattista and P. Chakraborty, “Quantitative characterization of the urine and serum metabolomes of children is essential for ‘omics’ studies,” *BMC Med.*, vol. 16, no. 1, p. 222, Dec. 2018.
[139] B. Muqaku *et al.*, “Multi-omics Analysis of Serum Samples Demonstrates Reprogramming of Organ Functions Via Systemic Calcium Mobilization and Platelet Activation in Metastatic Melanoma,” *Mol. Cell. Proteomics*, vol. 16, no. 1, pp. 86–99, Jan. 2017.
[140] P. Díez and M. Fuentes, “Proteogenomics for the Comprehensive Analysis of Human Cellular and Serum Antibody Repertoires,” 2016, pp. 153–162.
[141] R. Srivastava *et al.*, “Serum profiling of leptospirosis patients to investigate proteomic alterations,” *J. Proteomics*, vol. 76, pp. 56–68, Dec. 2012.
[142] E. Nagata, “Identification of biomarkers associated with migraine,” *Rinsho Shinkeigaku*, vol. 52, no. 11, pp. 1014–1017, 2012.
[143] A. P. Drabovich, P. Saraon, K. Jarvi, and E. P. Diamandis, “Seminal plasma as a diagnostic fluid for male reproductive system disorders,” *Nat. Rev. Urol.*, vol. 11, no. 5, pp. 278–288, May 2014.
[144] A. Meyer *et al.*, “Plasma metabolites and lipids predict insulin sensitivity improvement in obese, nondiabetic individuals after a 2-phase dietary intervention,” *Am. J. Clin. Nutr.*, vol. 108, no. 1, pp. 13–23, Jul. 2018.

[145] L. M. Heaney, K. Deighton, and T. Suzuki, “Non-targeted metabolomics in sport and exercise science,” *J. Sports Sci.*, vol. 37, no. 9, pp. 959–967, May 2019.
[146] “Metabolomics,” *Metabolomics*, vol. 15, no. 9, 2019.
[147] “OMICS,” *Omi. A J. Integr. Biol.*, vol. 23, no. 8, 2019.
[148] “Metabolics Society.” [Online]. Available: http://metabolomicssociety.org/. [Accessed: 27-Aug-2019].
[149] U. S. N. L. of M. NIH, “Metabolomics Workbench.” [Online]. Available: https://www.metabolomicsworkbench.org. [Accessed: 27-Aug-2019].
[150] B. Karahalil, “Overview of Systems Biology and Omics Technologies,” *Curr. Med. Chem.*, vol. 23, no. 37, pp. 4221–4230, Dec. 2016.
[151] C. Manzoni *et al.*, “Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences,” *Brief. Bioinform.*, vol. 19, no. 2, pp. 286–302, Mar. 2018.
[152] B. B. Misra, C. Langefeld, M. Olivier, and L. A. Cox, “Integrated omics: tools, advances and future approaches,” *J. Mol. Endocrinol.*, pp. R21–R45, Jan. 2019.
[153] Inverse, “Here’s Google ‘Quantum Supremacy’ paper it pulled from NASA’s website,” 2019. [Online]. Available: https://www.inverse.com/article/59507-full-quantum-supremacy-paper. [Accessed: 25-Sep-2019].
[154] D. I. Ellis, W. B. Dunn, J. L. Griffin, J. W. Allwood, and R. Goodacre, “Metabolic fingerprinting as a diagnostic tool,” *Pharmacogenomics*, vol. 8, no. 9, pp. 1243–1266, 2007.
[155] G. Theophilou, M. Paraskevaidi, K. M. Lima, M. Kyrgiou, P. L. Martin-Hirsch, and F. L. Martin, “Extracting biomarkers of commitment to cancer development: Potential role of vibrational spectroscopy in systems biology,” *Expert Rev. Mol. Diagn.*, vol. 15, no. 5, pp. 693–713, 2015.
[156] “The importance of metabolomics,” *European Bioinformatics Institute*. [Online]. Available: https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/importance-metabolomics. [Accessed: 27-Aug-2019].
[157] X. Wu, J. Zhu, B. Wu, J. Sun, and C. Dai, “Discrimination of tea varieties using FTIR spectroscopy and allied Gustafson-Kessel clustering,” *Comput. Electron. Agric.*, vol. 147, no. 301, pp. 64–69, Apr. 2018.
[158] Y. Fujimura *et al.*, “Metabolomics-driven nutraceutical evaluation of diverse green tea cultivars,” *PLoS One*, vol. 6, no. 8, 2011.
[159] Q. Zhang, M. Liu, and J. Ruan, “Metabolomics analysis reveals the metabolic and functional roles of flavonoids in light-sensitive tea leaves,” *BMC Plant Biol.*, vol. 17, no. 1, pp. 1–10, 2017.
[160] K. Navratilova *et al.*, “Green tea: Authentication of geographic origin based on UHPLC-HRMS fingerprints,” *J. Food Compos. Anal.*, vol. 78, pp. 121–128, May 2019.
[161] J. Liu, Q. Zhang, M. Liu, L. Ma, Y. Shi, and J. Ruan, “Metabolomic Analyses Reveal Distinct Change of Metabolites and Quality of Green Tea during the Short Duration of a Single Spring Season,” *J. Agric. Food Chem.*, vol. 64, no. 16, pp. 3302–3309, Apr. 2016.
[162] A. Derenne, V. Van Hemelryck, D. Lamoral-Theys, R. Kiss, and E. Goormaghtigh, “FTIR spectroscopy: A new valuable tool to classify the effects of polyphenolic compounds on cancer cells,” *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1832, no. 1, pp. 46–56, Jan. 2013.
[163] W. C. Reygaert, “Green Tea Catechins: Their Use in Treating and Preventing Infectious Diseases,” *Biomed Res. Int.*, vol. 2018, pp. 1–9, Jul. 2018.
[164] C. S. Robb, S. E. Geldart, J. A. Seelenbinder, and P. R. Brown, “ANALYSIS OF GREEN TEA CONSTITUENTS BY HPLC-FTIR,” *J. Liq. Chromatogr. Relat. Technol.*, vol. 25, no. 5, pp. 787–801, Apr. 2002.
[165] S. Sivakumar *et al.*, “FT-IR study of green tea leaves and their diseases of Arunachal Pradesh, North East, India,” *Pharm. Chem. J.*, vol. 1, no. 3, pp. 17–24, 2014.
[166] S. R. Senthilkumar and S. Thirumal, “Green tea (Camellia sinensis) mediated synthesis of zinc oxide (ZnO) nanoparticles and studies on their antimicrobial activities,” *Int. J. Pharm. Pharm. Sci.*, vol. 6, pp. 461–465, 2014.
[167] S. Agatonovic-Kustrin, “The Use of Fourier Transform Infrared (FTIR) Spectroscopy and Artificial Neural Networks (ANNs) to Assess Wine Quality,” *Mod. Chem. Appl.*, vol. 01, no. 04, 2013.
[168] Y.-T. Wang *et al.*, “FTIR spectroscopy coupled with machine learning approaches as a rapid

tool for identification and quantification of artificial sweeteners," *Food Chem.*, vol. 303, p. 125404, Jan. 2020.
[169] E. P. Mwanga *et al.*, "Using mid-infrared spectroscopy and supervised machine-learning to identify vertebrate blood meals in the malaria vector, Anopheles arabiensis," *Malar. J.*, vol. 18, no. 1, p. 187, 2019.
[170] M. G. Madden and T. Howley, "A Machine Learning Application for Classification of Chemical Spectra," in *Applications and Innovations in Intelligent Systems XVI*, London: Springer London, 2009, pp. 77–90.
[171] H. H. Mantsch, "The road to medical vibrational spectroscopy - A history," *Analyst*, vol. 138, no. 14, pp. 3863–3870, 2013.
[172] D. A. Skoog, F. J. Holler, and S. R. Crouch, *Principles of Instrumental Analysis*, Seventh. Brooks Cole, 2017.
[173] L. G. Wade, *Organic Chemistry*, Eight. Pearson, 2012.
[174] T. Higdon, "FT-IR Spectroscopy Technology, Market Evolution and Future Strategies of Bruker Optics Inc . by FT-IR Spectroscopy Technology, Market Evolution and Future Strategies of Bruker Optics Inc .," Massachusetts Institute of Technology, 2010.
[175] J. M. Chalmers and P. R. Griffiths, *Handbook of Vibrational Spectroscopy*. Chichester, UK: John Wiley & Sons, Ltd, 2001.
[176] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*. Chichester, UK: John Wiley & Sons, Ltd, 2004.
[177] M.-M. Blum and H. John, "Historical perspective and modern applications of Attenuated Total Reflectance - Fourier Transform Infrared Spectroscopy (ATR-FTIR)," *Drug Test. Anal.*, vol. 4, no. 3–4, pp. 298–302, Mar. 2012.
[178] Á. I. López-Lorente and B. Mizaikoff, "Mid-infrared spectroscopy for protein analysis: potential and challenges," *Anal. Bioanal. Chem.*, vol. 408, no. 11, pp. 2875–2889, Apr. 2016.
[179] K. Hauser, "Infrared Spectroscopy of Protein Folding, Misfolding and Aggregation," in *Encyclopedia of Biophysics*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1089–1095.
[180] J. Kong and S. Yu, "Fourier transform infrared spectroscopic analysis of protein secondary structures," *Acta Biochim. Biophys. Sin. (Shanghai).*, vol. 39, no. 8, pp. 549–559, 2007.
[181] R. K. Sahu *et al.*, "Characteristic Absorbance of Nucleic Acids in the Mid-IR Region as Possible Common Biomarkers for Diagnosis of Malignancy," *Technol. Cancer Res. Treat.*, vol. 3, no. 6, pp. 629–638, Dec. 2004.
[182] M. Banyay, M. Sarkar, and A. Gräslund, "A library of IR bands of nucleic acids in solution," *Biophys. Chem.*, vol. 104, no. 2, pp. 477–488, Jun. 2003.
[183] B. R. Wood, "The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues," *Chem. Soc. Rev.*, vol. 45, no. 7, pp. 1980–1998, 2016.
[184] M. J. Baker *et al.*, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nat. Protoc.*, vol. 9, no. 8, pp. 1771–1791, Aug. 2014.
[185] Z. Movasaghi, S. Rehman, and D. I. ur Rehman, "Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues," *Appl. Spectrosc. Rev.*, vol. 43, no. 2, pp. 134–179, Feb. 2008.
[186] S.-R. Tsai and M. R. Hamblin, "Biological effects and medical applications of infrared radiation," *J. Photochem. Photobiol. B Biol.*, vol. 170, pp. 197–207, May 2017.
[187] J. J. Workman, "Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999," *Appl. Spectrosc. Rev.*, vol. 34, no. 1–2, pp. 1–89, 1999.
[188] F. Toor, S. Jackson, X. Shang, S. Arafin, and H. Yang, "Mid-infrared Lasers for Medical Applications: introduction to the feature issue," *Biomed. Opt. Express*, vol. 9, no. 12, p. 6255, Dec. 2018.
[189] N. K. Niazi, B. Singh, and B. Minasny, "Mid-infrared spectroscopy and partial least-squares regression to estimate soil arsenic at a highly variable arsenic-contaminated site," *Int. J. Environ. Sci. Technol.*, vol. 12, no. 6, pp. 1965–1974, Jun. 2015.
[190] N. N. Misra, C. Sullivan, and P. J. Cullen, "Process Analytical Technology (PAT) and Multivariate Methods for Downstream Processes," *Curr. Biochem. Eng.*, vol. 2, no. 1, pp. 4–16, Apr. 2015.
[191] D. N. Stratis-Cullum *et al.*, "Spectroscopic data in biological and biomedical analysis,"

Biomed. Photonics Handbook, Second Ed. Fundam. Devices, Tech., pp. 587–800, 2014.
[192] B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*, Second. CRC Press, 2011.
[193] B. H. Stuart, "Infrared Spectroscopy of Biological Applications: An Overview," in *Encyclopedia of Analytical Chemistry*, Chichester, UK: John Wiley & Sons, Ltd, 2012.
[194] C. P. Schultz, "Precision infrared spectroscopic imaging: The future of FT-IR spectroscopy," *Spectroscopy*, vol. 16, no. October, pp. 24–33, 2001.
[195] K. Ataka, T. Kottke, and J. Heberle, "Thinner, smaller, faster: IR techniques to probe the functionality of biological and biomimetic systems," *Angew. Chemie - Int. Ed.*, vol. 49, no. 32, pp. 5416–5424, 2010.
[196] D. Perez-Guaita, S. Garrigues, and M. de la, "Infrared-based quantification of clinical parameters," *TrAC - Trends Anal. Chem.*, vol. 62, pp. 93–105, 2014.
[197] C. Hughes *et al.*, "Assessing the challenges of Fourier transform infrared spectroscopic analysis of blood serum," *J. Biophotonics*, vol. 7, no. 3–4, pp. 180–188, 2014.
[198] J. R. Mourant *et al.*, "Methods for measuring the infrared spectra of biological cells," *Phys. Med. Biol.*, vol. 48, no. 2, pp. 243–257, 2003.
[199] M. Wenning and S. Scherer, "Identification of microorganisms by FTIR spectroscopy: Perspectives and limitations of the method," *Appl. Microbiol. Biotechnol.*, vol. 97, no. 16, pp. 7111–7120, 2013.
[200] C. Hughes *et al.*, "SR-FTIR spectroscopy of renal epithelial carcinoma side population cells displaying stem cell-like characteristics," *Analyst*, vol. 135, no. 12, pp. 3133–3141, 2010.
[201] A. Sevinc, D. Yonar, and F. Severcan, "Investigation of neurodegenerative diseases from body fluid samples using Fourier transform infrared spectroscopy," *Biomed. Spectrosc. Imaging*, vol. 4, no. 4, pp. 341–357, 2015.
[202] L. M. Miller, M. W. Bourassa, and R. J. Smith, "FTIR spectroscopic imaging of protein aggregation in living cells," *Biochim. Biophys. Acta - Biomembr.*, vol. 1828, no. 10, pp. 2339–2346, 2013.
[203] G. Bellisola *et al.*, "Rapid recognition of drug-resistance/sensitivity in leukemic cells by fourier transform infrared microspectroscopy and unsupervised hierarchical cluster analysis," *Analyst*, vol. 138, no. 14, pp. 3934–3945, 2013.
[204] D. Landgrebe *et al.*, "On-line infrared spectroscopy for bioprocess monitoring," *Appl. Microbiol. Biotechnol.*, vol. 88, no. 1, pp. 11–22, 2010.
[205] J. G. Leite and J. T. Cavalheiro, "Aplicação das Técnicas de Espectroscopia FTIR e de Micro Espectroscopia Confocal Raman à Preservação do Património," no. 020805011, p. 76, 2008.
[206] K. Z. Liu, M. Xu, and D. A. Scott, "Biomolecular characterisation of leucocytes by infrared spectroscopy," *Br. J. Haematol.*, vol. 136, no. 5, pp. 713–722, 2007.
[207] J. L. Moreira, A. M. Marcos, and P. Barros, "Potencialidades da Espectrometria de Infravermelho por Transformada de Fourier (FTIR) na análise de vinhos," p. 4400, 2000.
[208] M. Heneczkowski, M. Kopacz, D. Nowak, and A. Kuzniar, "Infrared spectrum analysis of some flavonoids.," *Acta Pol. Pharm.*, vol. 58, pp. 415–420, 2001.
[209] Z. Guo, Q. Chen, L. Chen, W. Huang, C. Zhang, and C. Zhao, "Optimization of Informative Spectral Variables for the Quantification of EGCG in Green Tea Using Fourier Transform Near-Infrared (FT-NIR) Spectroscopy and Multivariate Calibration," *Appl. Spectrosc.*, vol. 65, no. 9, pp. 1062–1067, Sep. 2011.
[210] D. Perez-Guaita, J. Kuligowski, G. Quintás, S. Garrigues, and M. De La Guardia, "Atmospheric compensation in fourier transform infrared (FT-IR) spectra of clinical samples," *Appl. Spectrosc.*, vol. 67, no. 11, pp. 1339–1342, 2013.
[211] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemom. Intell. Lab. Syst.*, vol. 117, pp. 100–114, 2012.
[212] X. Shen *et al.*, "Study on baseline correction methods for the Fourier transform infrared spectra with different signal-to-noise ratios," *Appl. Opt.*, vol. 57, no. 20, p. 5794, Jul. 2018.
[213] L. M. Miller and P. Dumas, "Infrared Spectroscopy using Synchrotron Radiation," in *Encyclopedia of Biophysics*, vol. 50, no. 11, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1106–1112.
[214] J. Sedman, A. Ghetler, A. Enfield, and A. A. Ismail, "Infrared Imaging: Principles and Practices," *Handb. Vib. Spectrosc.*, pp. 1–23, 2010.

[215] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, “Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra,” *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, 1989.
[216] Å. Rinnan, F. van den Berg, and S. B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra,” *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009.
[217] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, “Review of multidimensional data processing approaches for Raman and infrared spectroscopy,” *EPJ Tech. Instrum.*, vol. 2, no. 1, 2015.
[218] P. R. Griffiths, “Introduction to the Theory and Instrumentation for Vibrational Spectroscopy,” *Handb. Vib. Spectrosc.*, 2010.
[219] F. Rosa *et al.*, “Monitoring the ex-vivo expansion of human mesenchymal stem/stromal cells in xeno-free microcarrier-based reactor systems by MIR spectroscopy,” *Biotechnol. Prog.*, vol. 32, no. 2, pp. 447–455, 2016.
[220] G. Vivó-Truyols and P. J. Schoenmakers, “Automatic selection of optimal Savitzky-Golay smoothing,” *Anal. Chem.*, vol. 78, no. 13, pp. 4598–4608, 2006.
[221] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.,” *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.
[222] K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, First. CRC Press, 2009.
[223] L. Hatcher, *Advanced Statistics in Research: Reading, Understanding, and Writing Up Data Analysis Results*. Shadow Finch Media LLC, 2013.
[224] K. H. Esbensen, B. Swarbrick, F. Westad, P. Whitcomb, and M. Anderson, *Multivariate Data Analysis: An introduction to Multivariate Analysis, Process Analytical Technology and Quality by Design*, Sixth. Oslo: CAMO Software AS, 2018.
[225] J. F. Hair, *Multivariate Data Analysis*, Seventh. PEL, 2013.
[226] K. H. Esbensen, D. Guyot, F. Westad, and L. P. Houmøller, “Multivariate Data Analysis - In Practice - An Introduction to Multivariate Data Analysis and Experimental Design,” no. 91, 2004.
[227] J. Shlens, “A Tutorial on Principal Component Analysis,” 2014.
[228] E. Bingham, S. Kaski, J. Laaksonen, and J. Lampinen, *Advances in Independent Component Analysis and Learning Machines*, First. Academic Press, 2015.
[229] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016.
[230] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.
[231] W. N. Haining, “The Numerology of T Cell Functional Diversity,” *Immunity*, vol. 36, no. 1, pp. 10–12, 2012.
[232] W. Lisik *et al.*, “Down regulation of genes involved in T cell polarity and motility during the induction of heart allograft tolerance by allochimeric MHC I,” *PLoS One*, vol. 4, no. 12, 2009.
[233] Y. Li, Z. Wu, J. Wei, A. Plaza, J. Li, and Z. Wei, “Fast principal component analysis for hyperspectral imaging based on cloud computing,” *Int. Geosci. Remote Sens. Symp.*, vol. 2015-Novem, pp. 513–516, 2015.
[234] M. Ernst, R. A. Dawud, A. Kurtz, G. Schotta, L. Taher, and G. Fuellen, “Comparative computational analysis of pluripotency in human and mouse stem cells,” *Sci. Rep.*, vol. 5, pp. 1–15, 2015.
[235] S. Do Cho, B. Hwan Hyun, and J. K. Kim, “Assessment of technological level of stem cell research using principal component analysis,” *Springerplus*, vol. 5, no. 1, 2016.
[236] P. K. Kimes, Y. Liu, D. Neil Hayes, and J. S. Marron, “Statistical significance for hierarchical clustering,” *Biometrics*, vol. 73, no. 3, pp. 811–821, 2017.
[237] A. K. Jain, M. N. Murty, P. J. Flynn, C. Methodologies, and I. Storage, “Data Clustering : A Review,” vol. 1, no. 212.
[238] M. K. Rafsanjani, Z. A. Varzaneh, and N. E. Chukanlo, “A survey of hierarchical clustering algorithms,” *Math. Comput. Sci.*, vol. 5, no. 3, pp. 229–240, 2012.
[239] C. Lima, L. Correa, H. Byrne, and D. Zezell, “K-means and Hierarchical Cluster Analysis as segmentation algorithms of FTIR hyperspectral images collected from cutaneous tissue,” in

2018 SBFoton International Optics and Photonics Conference (SBFoton IOPC), 2018, pp. 1–4.
[240] Q. Zhong *et al.*, "Similarity maps and hierarchical clustering for annotating FT-IR spectral images," *BMC Bioinformatics*, vol. 14, no. 1, p. 333, 2013.
[241] W. Sarada and P. V. Kumar, "A REVIEW ON CLUSTERING TECHNIQUES AND THEIR COMPARISON," *Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 11, pp. 2806–2812, 2013.
[242] K. Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review," *Int. J. Sci. Res. Publ.*, vol. 3, no. 3, pp. 2–4, 2013.
[243] C. Kirschner, N. A. Ngo Thi, and D. Naumann, "FT-IR spectroscopic investigations of antibiotic sensitive and resistant microorganisms," in *Spectroscopy of Biological Molecules: New Directions*, vol. 3257, no. 1991, Dordrecht: Springer Netherlands, 1999, pp. 561–562.
[244] D. Savić, N. Joković, and L. Topisirović, "Multivariate statistical methods for discrimination of lactobacilli based on their FTIR spectra," *Dairy Sci. Technol.*, vol. 88, no. 3, pp. 273–290, May 2008.
[245] R. Davis and L. J. Mauer, "Fourier tansform infrared (FT-IR) spectroscopy : A rapid tool for detection and analysis of foodborne pathogenic bacteria," no. 1, pp. 1582–1594, 2010.
[246] B. Zimmermann, "Characterization of Pollen by Vibrational Spectroscopy," vol. 64, no. 12, pp. 1364–1373, 2010.
[247] K. Ali *et al.*, "Fourier transform infrared spectromicroscopy and hierarchical cluster analysis of human meningiomas," *Int. J. Mol. Med.*, pp. 297–301, 2008.
[248] E. T. Jaynes, *Probability Theory, The Logic of Science*. 2003.
[249] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Third. Chapman and Hall/CRC, 2013.
[250] R. H. RIFFENBURGH, "Methods You Might Meet, But Not Every Day," in *Statistics in Medicine*, Elsevier, 2006, pp. 521–529.
[251] Q. Shuang, Y. Yuan, M. Zhang, and D. Yu, "Bankruptcy prediction in construction companies via Fisher's Linear Discriminant Analysis," in *2011 International Conference on E-Business and E-Government (ICEE)*, 2011, pp. 1–4.
[252] M. Chen, "Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches," *Comput. Math. with Appl.*, vol. 62, no. 12, pp. 4514–4524, 2011.
[253] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform," *Biomed. Signal Process. Control*, 2013.
[254] S. Rehman, F. Riaz, H. Ajmal, A. Hassan, Q. U. Ain, and A. Perwaiz, "PCA and LDA based classifiers for osteoporosis identification," in *Pattern Recognition and Tracking XXIX*, 2018, vol. 10649, p. 23.
[255] T. Jombart, S. Devillard, and F. Balloux, "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations," *BMC Genet.*, vol. 11, no. 1, p. 94, 2010.
[256] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 336–341.
[257] A. Khan and H. Farooq, "Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition," *Int. J. Comput. Sci. Issues (IJCSI)*, vol. 8, no. 6, pp. 267–270, 2011.
[258] M. Fordellone, A. Bellincontro, and F. Mencarelli, "Partial least squares discriminant analysis : A dimensionality reduction method to classify hyperspectral data," pp. 1–24, 2018.
[259] L. Tang, S. Peng, Y. Bi, P. Shan, and X. Hu, "A New Method Combining LDA and PLS for Dimension Reduction," *PLoS One*, vol. 9, no. 5, p. e96944, May 2014.
[260] S. Chevallier, D. Bertrand, A. Kohler, and P. Courcoux, "Application of PLS-DA in multivariate image analysis," *J. Chemom.*, vol. 20, no. 5, pp. 221–229, May 2006.
[261] A. Kalivodová, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, and T. Adam, "PLS-DA for compositional data with application to metabolomics," *J. Chemom.*, vol. 29, no. 1, pp. 21–28, Jan. 2015.
[262] L. C. Lee, C.-Y. Liong, and A. A. Jemain, "Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice

strategies and knowledge gaps," *Analyst*, vol. 143, no. 15, pp. 3526–3539, 2018.
[263] A. Biancolillo and F. Marini, "Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis," *Front. Chem.*, vol. 6, Nov. 2018.
[264] S. Sæbø, T. Almøy, A. Flatberg, A. H. Aastveit, and H. Martens, "LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables," *Chemom. Intell. Lab. Syst.*, vol. 91, no. 2, pp. 121–132, Apr. 2008.
[265] F. B. Lavoie, K. Muteki, and R. Gosselin, "Generalization of Powered–Partial-Least-Squares," *Chemom. Intell. Lab. Syst.*, vol. 179, pp. 1–11, Aug. 2018.
[266] E. Greenberg, "Minimum variance properties of principal component regression," *J. Am. Stat. Assoc.*, vol. 70, no. 349, pp. 194–197, 1975.
[267] R. C. Hill, "Multicollinearity and the value of a priori information," *Commun. Stat. - Theory Methods*, vol. 8, no. 5, pp. 477–486, 1979.
[268] M. G. Kendall, *A Course in Multivariate Analysis*. New York: Hafner Publishing Company, 1957.
[269] B. T. McCallum, "Artificial Orthogonalization in Regression Analysis," *Rev. Econ. Stat.*, vol. 52, no. 1, p. 110, Feb. 1970.
[270] C. S. AS, *The Unscrambler X user manual*. Oslo: CAMO Software AS, 2014.
[271] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.
[272] J. D. Kolsky, "Using Partial Least Squares Regression in Consumer Research," *Camo white Pap.*, pp. 1–4.
[273] K. Dunst, "Process Improvement using Data," 2019. [Online]. Available: https://learnche.org/pid/data-visualization/index. [Accessed: 24-Aug-2019].
[274] C. Ladeira, R. M., and E. Ribeiro, "Green Tea Epigallocatechin-3—gallate (EGCG) oxidative stress and DNA damage," in *47th Meeting of the European Environmental Mutagenesis & Genomics Society*, 2019.
[275] C. Ladeira, R. M., and E. Ribeiro, "Green Tea Epigallocatechin-3-gallate (EGCG) oxidative stress and DNA damage in vivo," *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.*
[276] "The Clinical and Biologic Evaluation of Polyphenon E, an Extract of Green Tea Containing EGCG, in Plasma Cell Dyscrasias - Pilot Study," *Clinical Trials*, 2015. [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT00942422. [Accessed: 28-Aug-2019].
[277] H. H. S. Chow *et al.*, "Pharmacokinetics and safety of green tea polyphenols after multiple-dose administration of epigallocatechin gallate and polyphenon E in healthy individuals," *Clin. Cancer Res.*, vol. 9, no. 9, pp. 3312–3319, 2003.
[278] F. S. Fan, "Iron deficiency anemia due to excessive green tea drinking," *Clin. Case Reports*, vol. 4, no. 11, pp. 1053–1056, 2016.
[279] N. A. Sachdev and M. Jothipriya, "Effect of Green Tea on Haemoglobin," *IOSR J. Dent. Med. Sci.*, vol. 16, no. 05, pp. 116–118, 2017.
[280] M. J. Baker *et al.*, "Developing and understanding biofluid vibrational spectroscopy: A critical review," *Chem. Soc. Rev.*, vol. 45, no. 7, pp. 1803–1818, 2016.
[281] F. Bonnier, M. J. Baker, and H. J. Byrne, "Vibrational spectroscopic analysis of body fluids: Avoiding molecular contamination using centrifugal filtration," *Anal. Methods*, vol. 6, no. 14, pp. 5155–5160, 2014.
[282] W. Tian, D. Wang, H. Fan, L. Yang, and G. Ma, "A plasma biochemical analysis of acute lead poisoning in a rat model by chemometrics-based fourier transform infrared spectroscopy: An exploratory study," *Front. Chem.*, vol. 6, no. JUN, pp. 1–8, 2018.
[283] E. Staniszewska-Slezak *et al.*, "Plasma biomarkers of pulmonary hypertension identified by Fourier transform infrared spectroscopy and principal component analysis," *Analyst*, vol. 140, no. 7, pp. 2273–2279, 2015.
[284] J. R. Hands *et al.*, "Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples," *J. Biophotonics*, vol. 7, no. 3–4, pp. 189–199, 2014.
[285] T. Y. Mahbubul Majumder, "Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing," *J. Data Mining Genomics Proteomics*, vol. 04, no. 04, 2013.
[286] C. Chen, Q. Liu, L. Liu, Y. Hu, and Q. Feng, "Potential Biological Effects of (-)-

Epigallocatechin-3-gallate on the Treatment of Nonalcoholic Fatty Liver Disease," *Mol. Nutr. Food Res.*, vol. 62, no. 1, p. 1700483, Jan. 2018.
[287] X. Li *et al.*, "Identification of Epigallocatechin-3- Gallate as an Inhibitor of Phosphoglycerate Mutase 1," *Front. Pharmacol.*, vol. 8, May 2017.
[288] M. P. Kapoor, M. Sugita, Y. Fukuzawa, and T. Okubo, "Physiological effects of epigallocatechin-3-gallate (EGCG) on energy expenditure for prospective fat oxidation in humans: A systematic review and meta-analysis," *J. Nutr. Biochem.*, vol. 43, pp. 1–10, May 2017.
[289] R.-Y. Gan, H.-B. Li, Z.-Q. Sui, and H. Corke, "Absorption, metabolism, anti-cancer effect and molecular targets of epigallocatechin gallate (EGCG): An updated review," *Crit. Rev. Food Sci. Nutr.*, vol. 58, no. 6, pp. 924–941, Apr. 2018.