TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

Detection of Caffeine Metabolites in Fingerprints by MALDI-MS Imaging

Ausgeführt am

Institut für Chemische Technologien und Analytik

der Technischen Universität Wien

Unter der Leitung von

Ao.Univ.Prof. Dr. **Johann Lohninger**

durch

Nada Eidi

Matr.Nr. 01529957

Kurzfassung

Spektroskopische Technologien werden seit einigen Jahrzehnten in Kombination mit hyperspektraler Bildgebung in einem breiten Anwendungsbereich für wissenschaftliche, klinische und kommerzielle Zwecke eingesetzt. Dazu gehören unter anderem Satellitenbildgebung, Qualitätskontrolle in der Lebensmittelindustrie und medizinische Diagnostik.

In letzter Zeit hat die Anwendung von Massenspektrometrie und hyperspektraler Bildgebung zur Analyse menschlicher Fingerabdrücke zugenommen. Viele Studien haben gezeigt, dass es möglich ist, aus einem individuellen Fingerabdruck schnell und nichtinvasiv detaillierte Informationen zu Lebensstil, Alter, Geschlecht und sogar zu Medikamenten- und Drogenkonsum zu erhalten.

In dieser Arbeit wird die Analyse menschlicher Fingerabdrücke unter Verwendung von MALDI MS (Matrix Assisted Laser Desorption/Ionization Mass Spectrometry) und hyperspektralen Bildgebungstechniken zum Erkennen des Koffeinverbrauchs und zur Bestimmung der Herkunftsregion einer Person diskutiert. Um dies zu erreichen, wurde ein Experiment mit sechs Freiwilligen aus zwei Ländern durchgeführt. Die Teilnehmer wurden gebeten, ihre Fingerabdrücke zweimal vor und nach dem Kaffeegenuss zu spenden.

Diese Studie demonstriert die einzelnen Schritte zur Durchführung und Analyse des Experiments, von der Erfassung der Fingerabdrücke über die Messung der Spektraldaten bis zur Interpretation der gewonnenen Daten. Die angewandten Auswertemethoden waren multiple lineare Regression, Hauptkomponentenregression und PLS/DA (Partial Least Squares Discriminant Analysis). Für jede Klassifizierungsmethode wurden zwei Modelle erstellt, um zwischen den Gruppen zu unterscheiden: Koffein / Nicht-Koffein und Personen aus verschiedenen Ländern.

Die Ergebnisse zeigten eine gute Klassifizierung der Gruppen. Allerdings konnte aufgrund der experimentellen Einschränkungen und der niedrigen Probenzahl, nicht statistisch signifikant nachgewiesen werden, dass diese Ergebnisse tatsächlichen einen Unterschied zwischen den Gruppen widerspiegeln. Es ist sehr wahrscheinlich, dass diese Ergebnisse durch Überanpassung generiert wurden. Diese Schlussfolgerung wird sowohl von den statistischen Kenngrößen der eingesetzten Verfahren als auch durch die Ergebnisse der Variablenauswahl unterstützt, die von etabliertem chemischem und spektroskopischem KnowHow abweichen. In der Arbeit werden Empfehlungen für eine Verbesserung des Verfahrens diskutiert.

Abstract

Spectrometry technologies combined with hyperspectral imaging have been used for decades in a wide range of applications for scientific, clinical and commercial uses. These include satellite imaging, food quality control, and pathogenic and diagnostic analysis, among others.

Recently, the utilization of mass spectrometry and hyperspectral imaging with human fingerprints has increased. Many studies showed that it is possible to obtain detailed information about lifestyle, age, gender, and even medication and drug consumption quickly and noninvasively from an individual fingerprint.

This study discusses analyzing human fingerprints using Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI MS) and hyperspectral imaging techniques to detect caffeine consumption and determine the region of origin of each person. To achieve this, an experiment with six volunteers from two countries was conducted. Participants were asked to donate their fingerprints twice, before and after coffee consumption.

This study demonstrates the experiment preparation steps from fingerprint acquisition to obtaining individual spectral data, preprocessing steps applied to overcome problems and issues raised during the experiment, and classification methods. The methods used were: multivariate regression, principal components regression, and partial least square discriminant analysis. For each classification method, two models were generated to differentiate between the groups: caffeine/non caffeine and individuals from different countries.

The results showed good classification of the groups, but due to experiment limitations, especially low sample number, it can't be proven that these results represent an actual difference between groups. It's highly likely these results have been generated randomly by overfitting all data points. This conclusion is supported by variable selection results which showed different variables from prior chemical knowledge, and by statistical tests results; therefore, some recommendations were discussed.

Acknowledgements

I would like to thank my advisor Ao.Univ.Prof. Dr. Johann Lohninger from the Institute of Chemical Technologies and Analytics at Vienna University of Technology (TU Wien), whose office door was always open whenever I ran into troubles or had a question. He consistently allowed this work to be my own work but steered me in the right direction whenever he thought I needed it.

I would also like to thank my colleagues who participated in the experiment for their patience during the preparation and measurements, without this help the work couldn't be done.

Special thanks, love and appreciation to my husband Adnan who always provided me with unconditional support and encouragement throughout my years of study and through preparing this thesis. This accomplishment would not have been possible without him.

Finally, I would like to thank my parents, siblings and my friends for their great support and love.

Thank you

List of abbreviations

9-AA	9-aminoacridine
Acetyl CoA	Acetyl coenzyme A
AMP	Adenosine monophosphate
AMS	Accelerator Mass Spectrometry
CF	Caffeine
CHCA	α-cyanohydroxycinnamic acid
CI	Chemical Ionization
COPD	Chronic Obstructive Pulmonary Diseases
DAN	Diaminonaphthalene
DESI-MS	Desorption Electrospray Ionization Mass Spectrometry
DHB	Dihydroxybenzoic acid
EI	Electron Ionization
ESI	Electrospray Ionization
GC	Gas-Chromatography
GC-MS	Gas Chromatography Mass Spectrometry
GMP	Guanosine monophosphate
HSI	Hyperspectral Imaging
ICP-MS	Inductively Coupled Plasma-MS
IMP	Inosine monophosphate
ITO	Indium Tin Oxide
LC	Liquid-Chromatography
LC-MS	Liquid Chromatography-MS
LDI-TOF MS	Laser Desorption/ Ionization Time of Flight Mass Spectrometry
LOD	Limit of Detection
MALDI	Matrix Assisted Laser Desorption Ionization
MLR	Multiple Linear Regression
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NIR	Near infrared
PCA	Principal Component Analysis
PCR	Principal Component Regression

PLS DA	Partial Least Squares Discriminant Analysis
Px	Paraxanthine
rDA	Retro-Diels–Alder reaction
SELDI-MS	Surface Enhanced Laser Desorption/Ionization-
SIMS	Secondary Ion Mass Spectrometry
SQ	Squalene
TAGs	Triacylglycerols
Tb	Theobromine
TLC	Thin-Layer Chromatography
TMS	Thermal infrared ranges
TN	True Negative
TOF	Time of Flight
Tp	Theophylline
TP	True Positive

List of Content

General introduction

Mass spectrometry technologies have been in use for many decades to conduct quantitative and qualitative studies, providing ability to analyze samples for physical, chemical, and biological purposes. This broad use came as the result of many distinctive features in comparison to other techniques like high sensitivity, low noise, and low damage of the samples of analysis, in addition to providing precise information of molecular weight and substances abundance. The integration of mass spectrometry and hyperspectral imaging technologies can provide further advantages and improve results, since they can combine spectra information with visual information of the samples. This can be used to determine the spatial disruption of compound of interest, meaning less noise, higher selectivity accuracy, and better input of statistical and mathematical algorithms that can be applied later. This combination has been used in many studies to achieve descriptive/predicative models.

Among many different fields, the analysis of vital signs in human bodies had always high potential. Many studies and applications using mass spectrometry and hyperspectral imaging have been introduced, including the analysis of human blood, saliva, urine, and fingerprint sweat. Among these substances, fingerprint samples have an advantage since their metabolites acquisition is a noninvasive process, can be conducted in a relatively short time, and can provide unique identification information when combined with ridge patterns.

This study attempts to provide a workflow description of the analysis of fingerprint metabolites using mass spectrometry/hyperspectral imaging space. All stages are described, beginning from fingerprint preparations and acquisitions and moving through molecular weight analysis steps, merging spectrum data with image data to form hypercube of information, and projecting the cube. Finally, the statistical analysis steps are discussed from data pre-processing to model generation.

The final aim for the case study provided in this thesis is to test the ability of the data provided by mass spectrometry and hyperspectral imaging to differentiate between caffeine/no caffeine consumption metabolic behaviors in the human body, and to differentiate between volunteers based on the region of origin from their fingerprints.

1 Fingerprints

1.1 Introduction

Fingerprints have been used in the forensic investigation field since the 19th century for the purpose of identifying individuals [1]. Nowadays, fingerprints are used for the detection of drugs, medication, and their metabolites, along with other biomolecules like lipids and proteins [2]. Generally, the individual's intake (whether by inhalation injection, digestion, etc.) of chemical or biological substances are deposited in the fingerprints due to sweat glands in the palms of hands and the ridges of the fingerprints [3]. The chemical components that can be found in the fingerprints, such as alcohol and nicotine, may reveal an individual's personal life style. Further personal information can be extracted from the finger's ridges and valleys, like gender and diet [4], by using mass spectrometry imaging techniques. These techniques clarify the relationship between the spatial distribution of the detected compound and its chemical information in one single analysis.

Many studies have reported various analytical methods for imaging the compounds which are deposited in the fingerprints, such as Desorption Electrospray Ionization Mass Spectrometry (DESI-MS) [5], Secondary Ion Mass Spectrometry (SIMS) [6], and Raman spectroscopy. Figure 1 shows the different biomolecules that are detected using Raman spectroscopy [7]. Among these techniques, Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI MS) has proven its ability to detect biological compounds like amino acids, peptides, proteins, and fatty acids [8], as well as chemical compounds like caffeine and its metabolites (theobromine, theophylline and paraxanthine) [9]. MALDI MS provides series of mass spectra for the fingerprint in one analysis and simultaneously includes high resolution and sensitivity [10].

In this chapter we will discuss the fingerprint types and their composition, including sweat and dissolved metabolites. Caffeine and lipid metabolism pathways in the human body are also described.

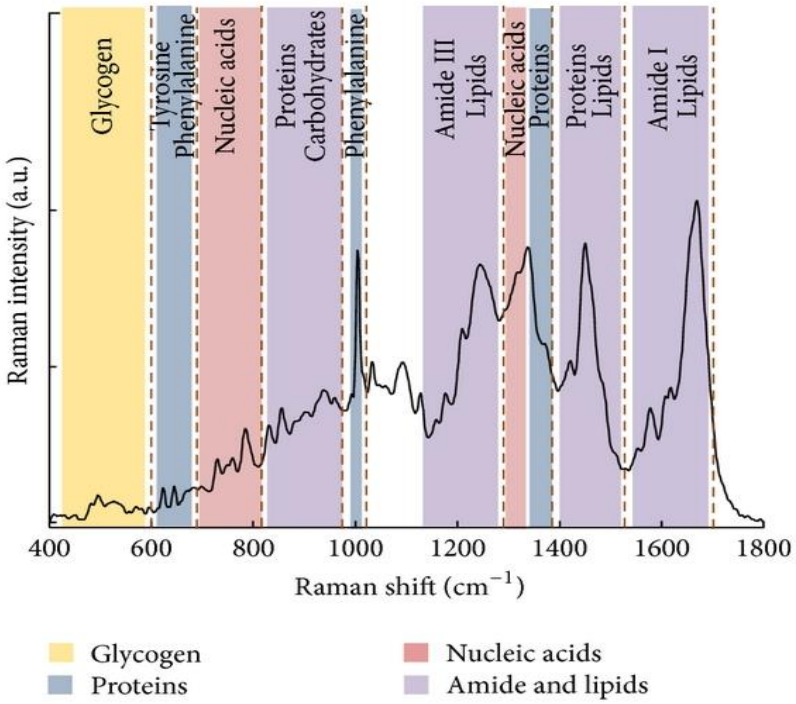


Figure 1: Scheme of the bands of Raman spectrum for different biomolecules present in fingerprints such as nucleic acids, proteins, lipids and many other molecules with different wavelengths [11].

1.2 Fingerprint composition and sweat glands

Fingerprints can be used for identification purposes as previously mentioned based on the fact that no two people have identical fingerprints, not even twins who share nearly identical physical characteristics [12]. However, this identification can't be performed for people who are prone to serious injury that affects the deep layer of their skin [13].

Fingerprints are classified into two categories. Invisible prints, also called latent prints, are invisible to the naked eye and form when a mixture of the body's natural oil and sweat from the skin are deposited onto a surface [14]. In other words, the fingerprints consist of different components that are originated from endogenous sources and transported to the fingerprint ridges. Visible prints, meanwhile, are formed when exogenous sources like blood, ink, dirt, food products, and cosmetics are transferred from the thumb or finger to the surface [15][16].

In latent prints, sweat that forms the major source of fingerprint composition originates from eccrine, sebaceous, and apocrine glands. Eccrine glands are found in the greatest density on the soles of the feet and palm of the hands. Sweat from this glands is the main contributor of the chemical components found on the fingerprint.

Sebaceous glands are found mostly in facial regions which contain hair follicles. These glands secrete an oily substance called sebum that includes cholesterol, free fatty acids, wax esters, and triglycerides. When a person touches their face during daily behaviors, sebum is transported from the face to the finger. The apocrine glands are found in the genital regions and the armpits, with any resulting sweat transported in the same way as the sebum to the finger [17].

A third category of molecules present in the fingerprint are called semi-endogenous substances. These substances result from the inhalation or ingestion of components such as drugs and medication, as well as food and drinks [18]. In other words, exogenous and semi-endogenous molecules enable the reconstruction of the individual's lifestyle and activities, while endogenous molecules give personal information about the individuals themselves such as the gender, diet, and medical conditions.

The personal information of an individual can be estimated from their fingerprints' composition and ridge density. Some studies noticed a significant difference in the fingerprints of males and females [19]. Specifically, ridges in male fingerprints are rougher and fattier than female fingerprints. Another difference between male and female fingerprints is that the concentrations of components like palmitic acid, palmitoleic acid, and oleic acid are higher in male fingerprints than female ones as determined by Gas Chromatography Mass Spectrometry (GC-MS) analysis [20].

The age of an individual also affects fingerprint composition because the amount and structure of lipids and fatty acids in the skin changes overtime. Ridges become more coarse as an individual ages [21].

1.3 Lipids in fingerprints

1.3.1 Introduction

The wide use of fingerprints in various fields arises from the fact that they contain a variety of chemical compounds, including inorganic substances like sodium, phosphate, and ammonia, as well as organic compounds like proteins, lipids, and amino acids. Variations of lipids and fatty acids can give personal information about an individual's age or gender, as previously mentioned. Moreover, lipids and fatty acids may also help determine the region of origin of an individual.

Lipids are one of the major biological molecules in the human body. They are involved in many processes like essential metabolic pathways, energy storage, cell signaling, and other functions like fat-soluble vitamins [22].

Many studies have found the principle role of lipids in the cell signaling systems. Certain lipids can act as cellular messengers [23], regulating many important functions such as cell growth, calcium mobilization, and programmed cell death [24][25].

Fat-soluble vitamins are usually stored in the liver and fatty tissues and are involved in many bioprocesses including substance concentration regulation, immune response activation, and neural functionalities [26].

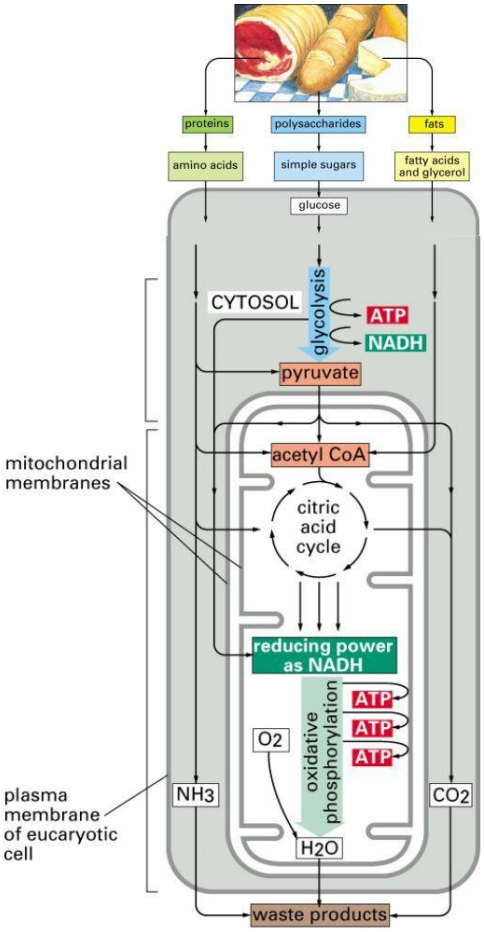


Figure 2: Energy pathways in the human body, where all intake types (proteins, carbohydrates and fats) are converted directly or indirectly to acetyl CoA, which is in turn converted to the basic energy unit ATP [27].

The human body can use different types of lipids to store energy, but triglycerides are still the main source of energy, which are usually stored in the adipose tissues. These tissues contain fat cells which work on the degradation and storage of triglycerides. Triglycerides come from two sources: intake fats including oil, butter and other dietary forms, and the conversion of excess carbohydrates into lipids by passing through many intermediate components.

All proteins, carbohydrates, and fats undergo a series of processes called lipo-genesis to produce acetyl coenzyme A (acetyl CoA) which is considered a crossroads for triglycerides synthesis as shown in Figure 2. Acetyl CoA is an intermediate compound that is involved in fatty acids storage in triacylglycerol form (shown in Figure 3-Right) as well as other forms of lipids such as cholesterol, cholesterol-derived steroid hormones, and sphingolipids, which are involved in signal transmission. The special crossroad role of acetyl-CoA is shown in Figure 3-Left.

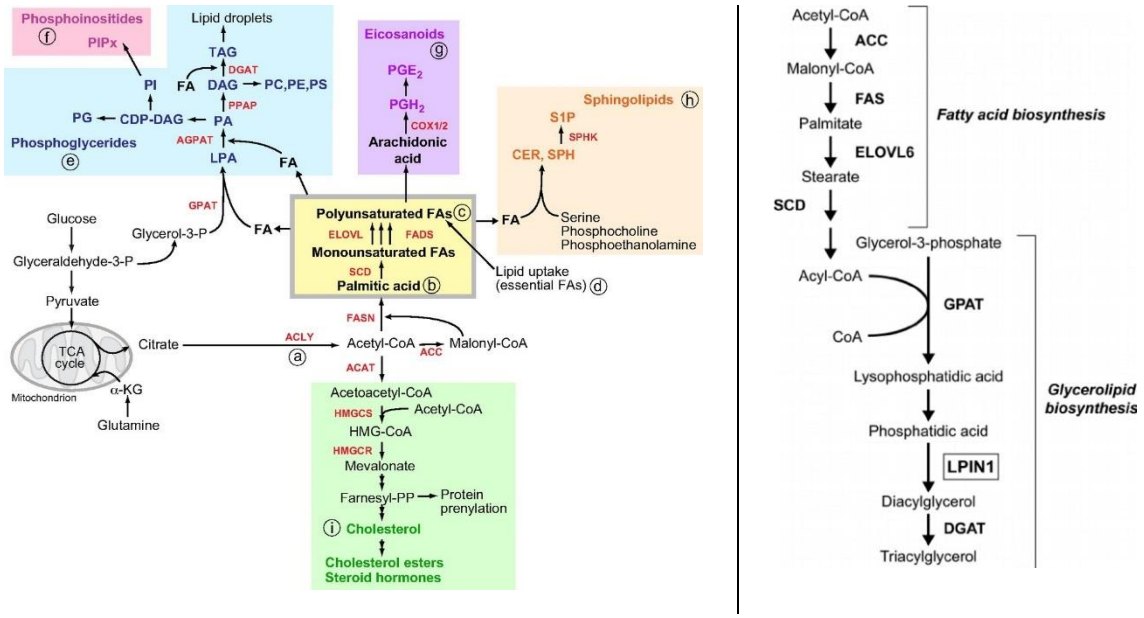


Figure 3 ***Left****: Schematic overview of the pathways involved in the synthesis of fatty acids (FAs), cholesterol, phosphoglycerides, eicosanoids and sphingolipids* [28]. ***Right****: Pathways of fatty acids and triglyceride syntheses from acetyl-CoA* [29].

The lipids degradation process is often initiated by lipase enzymes [30]. Lipases are a family of enzymes that catalyze the degradation of lipids by hydrolysis [31]. They contain many different types which are secreted by different tissues and cells to address each different lipid type's degradation processes. We recognize that PL, PLRP2, and PNLIP are secreted by pancreas [32]. LIPC is secreted by the liver [33], while LIPG and LPL are secreted by endothelial tissues [34] and other minor groups. On the contrary, some lipases are produced by pathogenic organisms, which can provide traces of some pathogens in the mucous and skin tissue [35]. This presents the possibility in the future of determining some pathogens directly from fingerprints if the technologies developed allow these kind of measurements.

1.3.2 Variation of lipids and fatty acids in the fingerprint

Because lipids and fatty acids determine many functions in the human body, they can provide information to detect an individual's characteristics. Lipids are secreted from the inner glands via metabolic pathways to the dermis layer of the fingerprint. Specifically, the sebaceous glands that primarily secret the sebum containing squalene (10%), triglycerols (25%), wax esters (22%), and free fatty acids (25%) as lipids mixture [36]. Moreover, the epidermal lipids are also involved in fingermark residues such as cholesterol (20%), fatty acids (65%), and ceramides [37]. These compounds have recently been detected using advanced analytical techniques. For example, triacylglycerols (TAGs) in the fingermark are analyzed as a biomarker of diseases using the Laser Desorption/ Ionization Time of Flight Mass Spectrometry (LDI-TOF MS) [38]. Squalene (SQ) is an organic compound which is extracted from the shark liver oil and is a biochemical precursor of the steroid family [39], it has a role in topical skin lubrication and protection and is involved in therapeutic drugs. SQ is well identified and separated in the fingerprint by using Thin-Layer Chromatography (TLC) and Electrospray Ionization (ESI)

techniques [40]. Fingerprints also contain cholesterol, which comes through the blood circulation and originates from the epidermal layer. It is involved in many fundamental processes within the cell and can also be a biomarker for some heart diseases and can be detected using laser desorption ionization Mass spectrometry (LDI-MS) [41].

In the previous discussion, we conclude that the variation of lipids consumption can be traced in the fingerprints. According to a recent study, it was possible to trace not only lipids but also other diet components [42]. Figure 4 shows how it is possible to trace six widespread plant oils in human fingerprints using mass spectrometry technologies.

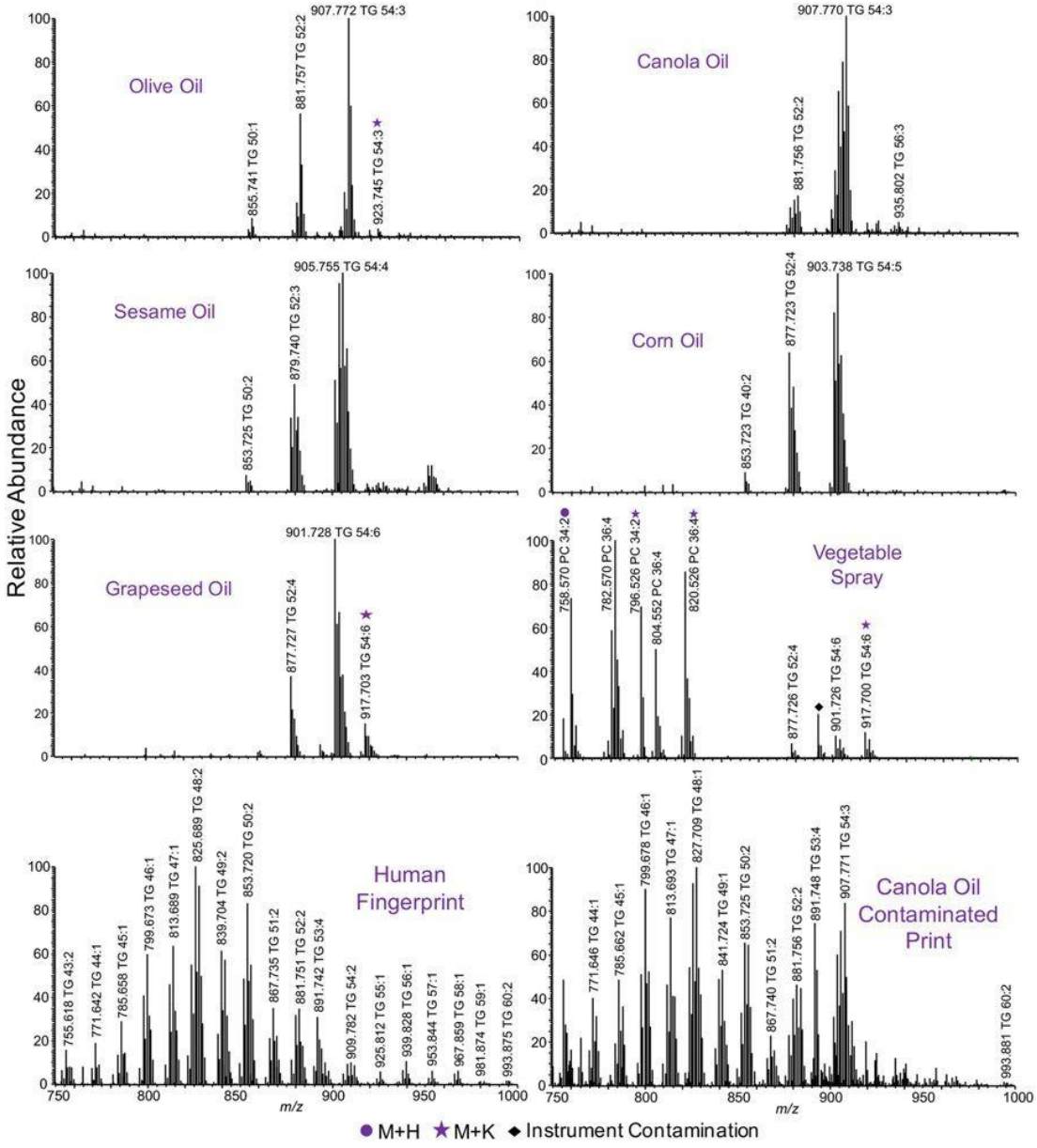


Figure 4: Positive mass spectrometry of six cooking oils (olive, canola, sesame, corn, grape seeds, and vegetable spray) with their fragmentations and relative abundances as well as their presence in human fingerprints [43].

1.4 Caffeine in fingerprints

1.4.1 Introduction

Caffeine (1,3,7-trimethylxanthine) is considered one of the most widely consumed psychoactive drugs and the most important naturally occurring xanthine alkaloid in the world [44]. According to Chinese legend, caffeine was first discovered by the Chinese emperor around 3000 BCE [45]. Caffeine (CF) is the basic element of coffee, tea, several energy drinks, and cola. Normally, caffeine's half-life in adults is about three to seven hours and can differ depending on factors such as the gender and age. Usually, the absorption time of CF from the digestive tract is about 45 minutes.

After intake, CF is metabolized in the electron transfer chain of the cell by hemeprotein called cytochrome p450 oxidase enzyme found in most human tissues, especially in the liver cells. This produces dimethylxanthines which in turn leads to form three primary isomeric metabolites for CF. An equivalent 98% of the CF ingestion is metabolized to paraxanthine (Px), theobromine (Tb), and theophylline (TP), while less than 2% of CF is excreted unchanged in human urine [46]. In this work, caffeine (CF) and its primary metabolites are the target analytes of detection in the fingerprint.

Caffeine comes in a white powder form, has a bitter taste, and is soluble in organic solvents and water with a percentage of 2.17% at room temperature. It has a molecular weight of approximately 194.19 g/mol with a boiling point at 178°C and melting point at 238°C. Caffeine also has a pH value of 6.9, which is considered as neutral [47]. The chemical structure of caffeine is shown in Figure 5: Caffeine chemical structure contains two fused rings, a pyrimidinedione, and imidazole .

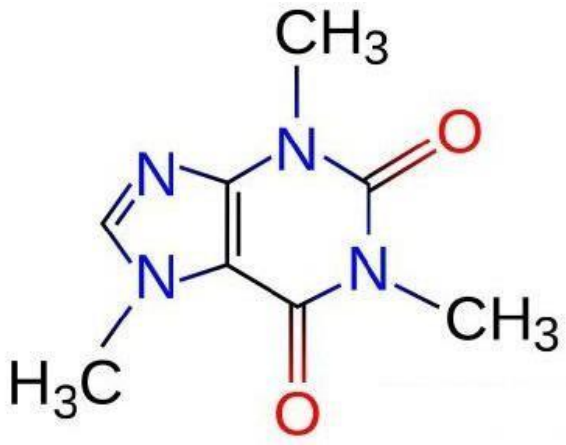


Figure 5: Caffeine chemical structure contains two fused rings, a pyrimidinedione, and imidazole [48].

The synthesis of caffeine in plants starts from IMP (Inosine monophosphate), AMP (Adenosine monophosphate), and GMP (Guanosine monophosphate), which are purine nucleotides that are turned into caffeine precursors (theobromine and theophylline) and then into caffeine by following different pathways (illustrated in Figure 6).

Generally, beverages that contain caffeine are consumed to improve human performance. Because CF is a central nervous system stimulator that reduces fatigue and drowsiness, it has effects on learning and memory skills and can improve reaction time, concentration, and motor coordination. These effects vary from person to person according to many factors such as body size and the degree of tolerance, in addition to medications, age, pregnancy, liver function, and enzymes that may influence caffeine absorption [49]. Like many central nervous system stimulants, an overdose of

caffeine (approximately 250-300 mg, in around 2-3 cups of coffee or 5-8 cups of tea per day) may lead to serious mental and physical symptoms such as irritability, headaches, nervousness, fidgeting, anxiety, gastrointestinal disturbance, and rapid heartbeat. These symptoms are accompanied with caffeinism [50]. Moreover, many dangerous symptoms can appear after a massive caffeine overdose such as depression, delusion, breakdown of skeletal muscle tissues, and can even cause death after around 200 mg per kg of body mass of CF intake [51]. For individuals who have chronic liver disease or a genetic disorder, however, the lethal dose can be lower than healthy individuals [52]. The various effects of caffeine on the human body, whether inhibitory or stimulatory, are shown in Figure 7.

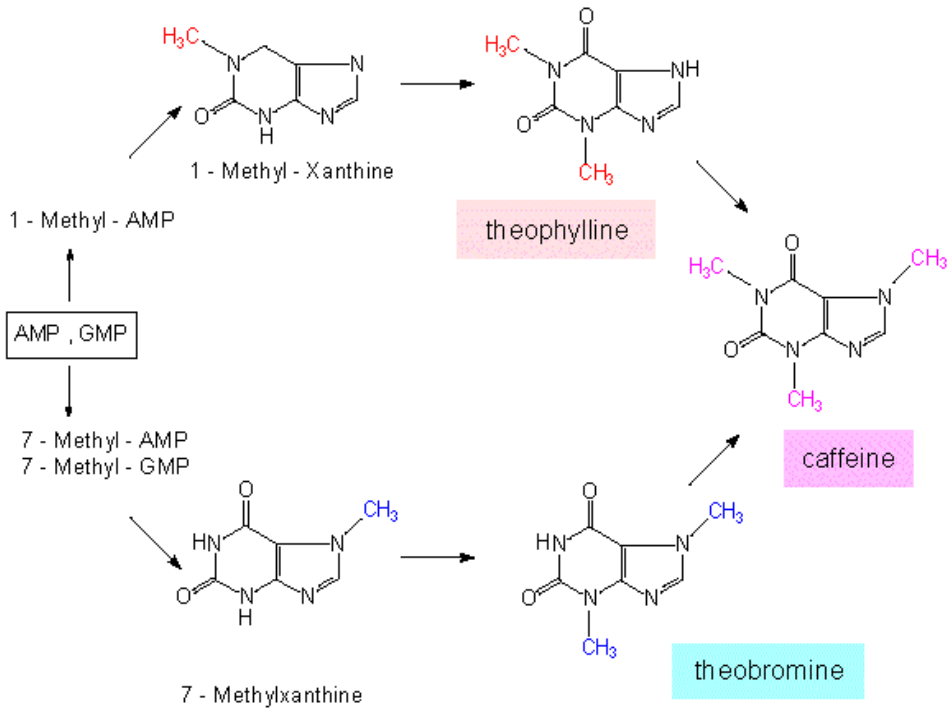


Figure 6: Caffeine synthesis in plants with two different pathways. The first path starts with 1-Methyl-AMP, which is transferred to theophylline, and another path starts with 7- Methyl- GMP and 7- Methyl- AMP, which are transferred into 7-Methylxanthine and then to theobromine. These compounds are the precursors of caffeine [53].

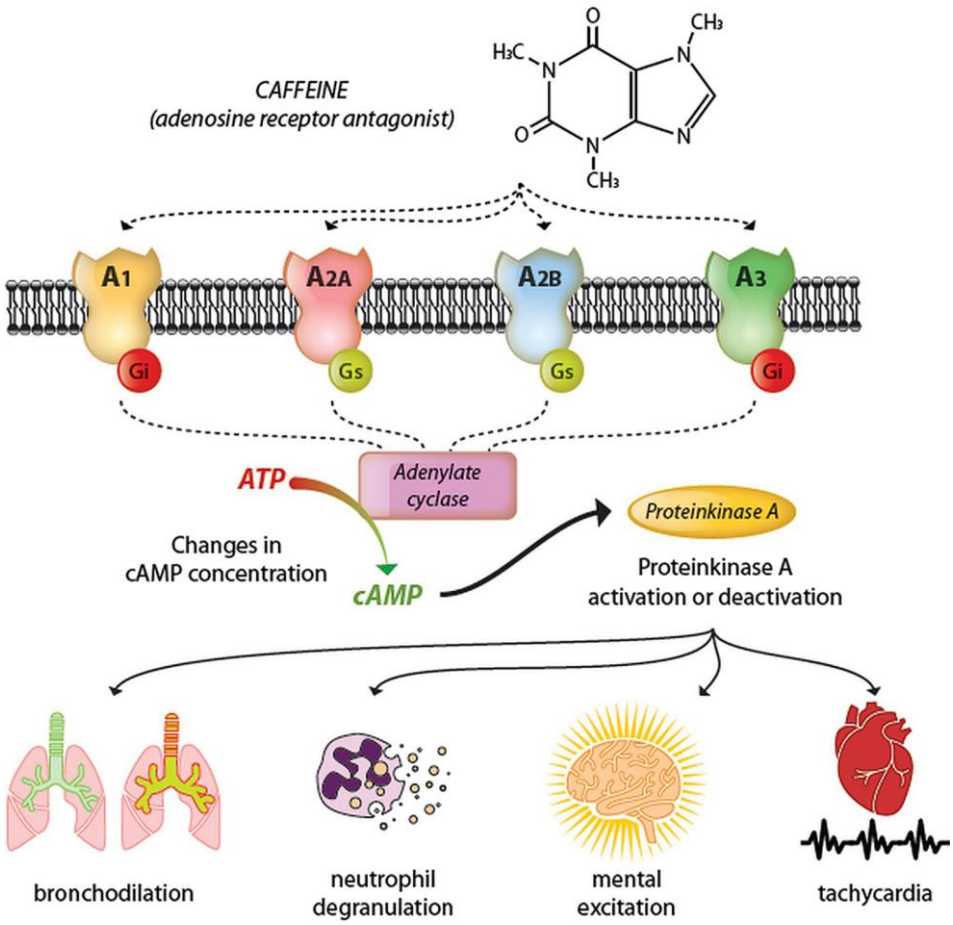


Figure 7: The various effects of caffeine on the functionality of human body organs can be divided into inhibitory or stimulatory according to interaction with different adenosine receptors. Protein kinase A enzyme can activate or deactivate some body functions and depend on the cellular level of cAMP. cAMP concentration is affected by caffeine absorption [54].

1.4.2 Caffeine primary metabolites

Paraxanthine PX (1,7-dimethylxanthine) is a central nervous stimulant with similar activity compared to caffeine. It forms about 84% of primary degradation of caffeine after metabolism in the liver by enzyme P450, while theobromine and theophylline form about 12% and 4% respectively of the CF degradation [55]. However, PX is the natural metabolite of caffeine in animals and some species like bacteria and is less toxic, showing less anxiogenic effect. It is generally not produced by plants [56].

Theobromine TB (3,7-dimethylxanthine) is a bitter alkaloid and the dominant methylxanthine found in chocolate, tea trees, kola nut, and other food. Although being the least effective of all the primary metabolites, it shows similar effects compared to caffeine but has less impact on the nervous system [57]. In the human body, the half-life of theobromine is about seven to twelve hours after consumption [58].

Theophylline TP (1,3- dimethylxanthine) is present in tea and cocoa and has similar structure and pharmacological effect in comparison with other methylxanthine derivatives. Theophylline is used for drug respiratory diseases such as Chronic Obstructive Pulmonary Diseases (COPD) and asthma due to its role in relaxing the smooth muscles in the bronchi [59]. However, it is also important to control the therapeutic dose of theophylline to avoid any toxicity and get the drug's useful therapeutic properties [60]. TP has a half-life between five and eight hours. The percentages and chemical structures of the caffeine metabolites are shown in Figure 8.

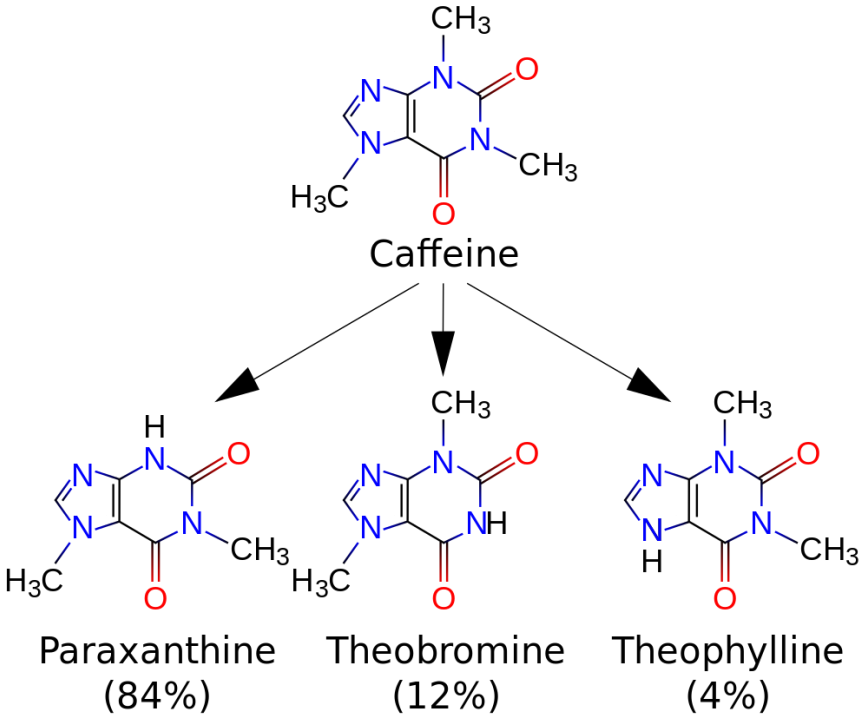


Figure 8: Caffeine and its primary metabolites, which show the percentage of metabolites in caffeine respectively (PX, TB, TP). All of these metabolites have the same chemical components and number of atoms, but with different arrangements [61].

2 Mass spectrometry

2.1 Introduction

In the past 30 years, mass spectrometry (MS) has undergone big developments in terms of both technical innovation and the extent of its application. Characteristics such as the unique sensitivity, detection limits, speed, and diversity of MS applications made it more dominant among other analytical methods.

MS is applied in many fields and can detect and identify pure and mixed substances in different phases (solid, liquid, gas). MS is an analytical tool that ionizes particles and molecules, breaking them into small charged fragments and classifying the generated products based on their mass to charge ratio. Qualitative and quantitative information of the sample can be obtained by applying various types of MS coupled with chromatographic techniques. Large and small biomolecules, as well as volatile and non-volatile compounds, can be detected with varying accuracy by MS such as peptides, organic compounds, lipids, and oligonucleotides.

MS was developed by a few dedicated proponents over many years. Between 1912 and 1913, the English physicist Joseph John Thomson used an immature form of MS called parabola spectrograph to separate particles of different mass to charge ratios. He also separated the ^{20}Ne and the ^{22}Ne isotopes, and correctly identified the m/z= 11 signal as a double charged ^{22}Ne particle [62]. However, the first fully functional mass spectrometer was built by Francis William Aston in 1919 in Cambridge. When he proved that some natural occurring elements like chlorine, bromine, and krypton consist of a combination of isotopes, he was awarded a Nobel Prize in 1922 after many years of experiments and studies developing MS. MS is proven to analyze not only chemical components, but also macromolecules using soft desorption /ionization methods. The latter procedure was invented by John Bennett Fenn and Koichi Tanaka who were awarded a Noble Prize in chemistry in 2002 for “the development of soft desorption ionization methods for mass spectrometric analyses of biological macromolecules” research.

2.2 Mass spectrometer principles

Mass spectrometers can detect mixed components in one analysis. This feature comes from coupling MS with separation techniques such as Gas-Chromatography (GC), Liquid-Chromatography (LC), and Thin Layer Chromatography (TLC), which work on separating the complex mixture over time by dissolving it in a fluid called mobile phase and passing it through another material called stationary phase. The components are then separated based on different traveling speeds. Each separation method is appropriate for specific components. For example, GC-MS is suitable for detecting stable substances up to 300°C. After separation, the resulting products are introduced into MS to be analyzed. Benefits from coupling MS with separation techniques include improving the selectivity, obtaining spectra for unknown and isolated components, and minimizing the products’ interference [63]. The main mass spectrometer block can be divided into fundamental parts: the ion source, mass analyzer, and the ion detector system, as shown in Figure 9.

Ion source: After introducing a sample into the inlet system, the neutral molecules of the sample enter the ionization chamber and are positively or negatively ionized and converted to gaseous phase. Generally, there are two types of ionization: hard ionization, which causes fragmentation of the sample molecules, and soft ionization, which analyzes the sample with minimum fragmentation.

Various types of ion sources are used according to the sample properties and the transferred energy through ionization processes. For example, in hard ionization, Chemical Ionization (CI) and Electron Ionization (EI) are suitable for gas-phase compounds and are usually coupled to Gas chromatography. Soft ionization methods such as Electrospray Ionization (ESI) and Matrix Assistance Laser Desorption /Ionization (MALDI) are often used for liquid and solid phase of the sample and are favorable for non-volatile substances. Generally, ionization methods play an essential role in mass spectrometry resolution. In hard ionization, highly detailed information can be obtained due to the high degree of fragmentation, but at the same time destroying the sample. Soft ionization keeps the substances intact with minimal fragmentation with achieving a high-resolution image.

Mass analyzer: The resulting ions are accelerated in a high vacuum chamber to prevent any collision while entering the flight path and are deflected by static or dynamic electric and magnetic fields. They are then separated according to their mass to charge ratio. For example, the molecular ions which have the same mass and kinetic energy will reach the detector at the same time, while the molecular ions with different weights and kinetic energies will reach the detector at different times. There are many types of mass analyzers, but they can generally be divided into three groups. The magnetic sector instruments allow the ions to enter the flight tube and be deflected by the magnetic field. The amount of deflection depends on the number of the charges and which quadrupoles are in the instrument. Another type of mass analyzer has been used for reflection enhancement, called Time of Flight (TOF), that is suitable for pulsed ion sources and works by measuring the flight time needed for the ions to reach the detector (see Chapter 4 for more details). The third mass analyzer group is called Ion-Trapp mass analyzers and includes linear ion traps and orbit traps. This group uses an oscillating electric field to store ions.

Detector system: The separated ions hit the detector and the steam of ions are amplified and sent to the computer to display the resulting signal as a graph, called the mass spectrum. The molecular ions with different m/z ratios are presented as m/z peaks in the x-axis and their intensities in the y-axis of the mass spectrum [64]. Furthermore, some data can be shown in a three-dimensional graph where the third axis (z-axis) records extra parameters like time. Ion detectors can be categorized in two groups: point ion collectors, which work on detecting ions one by one at a single point, and array collectors, which work on detecting all ions simultaneously along a plane. Several detectors are used depending on the experiment and the instrument. For example, a Secondary Electron Multiplier (SEM) is one type of point ion detector that can produce electrons and amplify them. This method increases the sensitivity and reduces noise and can be used to achieve a highly detailed image. Additional detectors are explained later in this work such as liner and reflector detectors, which are used mostly in MALDI MS (see Chapter 4).

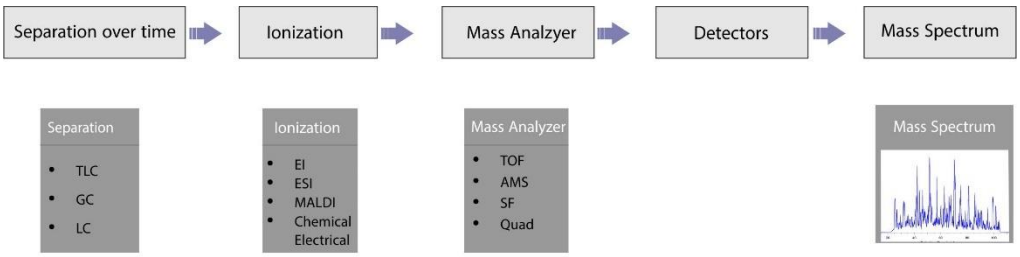


Figure 9: The mass spectrometer parts. Compounds are separated over time and enter the sample into the instrument, converting the sample particles to a gaseous phase in the ionization chamber. Ions are then sorted in a mass analyzer based on their m/z ratio and detected in the ion detection part. The resulting data is shown as a mass spectrum.

2.3 Types of mass spectrometers

As mentioned previously, MS can be coupled with several separation devices in order to identify a complex sample and any unknown components based on different chemical properties. This combination increases the accuracy of the identification process and reduces any potential error since some resulted fragments have similar mass spectra patterns. After this separation, MS can analyze the separated molecules precisely.

In this chapter, we will quickly overview the different types of MS and MS coupled with separation devices and their applications.

1. **Accelerator Mass Spectrometry (AMS)**: AMS works by applying high kinetic energies to ions and accelerating them before the mass analysis. It has a huge ability to isolate and detect rare and long-lived isotopes from other adjacent masses like (^{14}C, ^{36}CL and ^{26}AL). This strong property make it a widely used instrument in geophysics sciences and biomedical field [65].
2. **Gas Chromatography-MS (GC-MS)**: GC-MS consists of two blocks. In the GC block, a capillary column and an inert gas in a mobile phase such as helium is used. The second block is a normal MS chamber that analyzes the separated components. Different molecules are detected based on their features, and the molecules are separated into columns based on their chemical properties. The molecules then enter the MS block in order to be ionized, detected, and eventually form the mass spectrum. GC- MS is widely used in volatile molecules detection, drug detection, and fire investigation [66].
3. **Liquid Chromatography-MS (LC-MS)**: The traditional LC-MS is analogous to GC-MS and only differs in terms of its mobile phase, which here is a liquid solvent mixed with the required mixture. The separated components must enter an interface before the MS to avoid any incompatibility between them. In High-Performance Liquid Chromatography (HPLC), a high pressure (up to 350 bar) is applied in order to pass the mobile phase and uses a high-pressure liquid. LC-MS is suitable for identifying non-volatile compounds and is an important tool in biochemistry field [67].
4. **Inductively Coupled Plasma-MS (ICP-MS)**: ICP-MS is involved in various fields and is important in detecting a wild range of elements, metals, and their traces in ultra-low concentration [68]. In the sample preparation step, many separation devices such as LC or GC can be applied. The sample is mixed with an argon gas to form aerosol and carried to the plasma torch, which ionizes to produce ions. They then enter MS with a high-vacuum analyzer, leading these ions to pass through the aperture where they are detected and the mass spectra are observed.

5. **Matrix Assisted Laser Desorption/Ionization- MS (MALDI-MS)**: MALDI is a soft ionization method is used to analyze macromolecules as well as small molecules. The sample is irradiating with a pulsed laser by mixing it with nonvolatile, organic, UV sensitive compounds called matrix. After irradiating the sample and matrix with laser, the molecular ions can be formed and accelerated through the mass analyzer and the mass spectra are obtained. MALDI is a powerful instrument which is used in forensic investigations and especially in latent fingerprint detection [69].
6. **Surface Enhanced Laser Desorption/Ionization-MS (SELDI-MS)**. SELDI is a modified version of MALDI and it is appropriate for detecting proteins with low molecular weight. It follows the same procedure of MALDI but the only difference here is that the matrix plate has a protein binding characteristic that works on chromatographic separation of proteins. By that the strongest binding proteins stay in the plate while the weak ones are removed, and the rest steps are as MALDI manner. SELDI is used in proteomic studies in addition to cancers detection [70].
7. **Tandem Mass Spectrometry (MS/MS)**. MS/MS consists of several stages which the fragmentation process occurs between them by using multiple quadrupoles. Firstly, the ions are formed in the ion source using many methods like MALDI and ESI. Then, the resulted ions with a specific m/z ratio go through quadrupole to form fragments, these fragments pass to another quadrupole to form product ions and then they are detected using conventional detectors. It can be coupled to many analytical instruments and is broadly used for food contaminants detection [71].

2.4 Spectroscopy for caffeine and its metabolites

Since caffeine is the most abundant ingredient found in some beverages, many spectroscopic methodologies were used for CF detection. Generally, mass spectrometry presents the target molecules in a plot where the x-axis shows the m/z values of the molecular ions and the y-axis shows their intensities. However, we first need to understand how the mass spectrometer fragments the target molecules and creates the molecular ions. MS contains an ionization chamber that creates ions by using different methods such as electron ionization, chemical ionization, electrospray ionization, or Matrix-Assisted Laser Desorption/Ionization (MALDI) to minimize the fragmentation. The general fragmentation process begins from the following equation:

$$M^{\bullet\pm} \rightarrow X^{\pm} + Y^{\bullet}$$

This equation represents the simplest version of fragmentation patterns in MS. During the ionization step the molecules are given a certain amount of energy, making them exited and ionized ($M^{\bullet\pm}$). The molecular ion then breaks into two parts, turning into a positive or negative ion (x^+ or x^-) and an uncharged free radical ($Y^{\cdot}$). The formed ions will be accelerated, deflected, and detected to form the mass spectrum and the free radical will be neglected, meaning the spectrum peaks represent different fragments of the compound. Many MS techniques were able to detect caffeine and its metabolites in urine, blood, and saliva [72]. Furthermore, MALDI MS has proven its efficiency in identifying and detecting CF and its degradation compounds in latent fingerprints with high resolution images.

One aim of this work is to find a significant difference between the two groups of consumers (caffeine and non-caffeine). Thus, caffeine (1,3,7-trimethylxanthine) which has the chemical formula $C_8H_{10}N_4O_2$ and a mass spectrum with a fundamental peak at 194m/z, is our major analyte of interest. The fragmentation step and mass spectrum of CF begins by breaking down the intact CF molecule through the ionization process and forming a daughter ion at 165 m/z by losing one of the CHO functional group from CF molecule. Meanwhile, the 137 m/z peak is formed by elimination of methyl isocyanate CH_3-N=C=O by retro-Diels–Alder reaction (rDA) reaction [73]. Any additional loss of the carboxyl group CO leaves behind a fragment at 109 m/z. The sequential peak at 82m/z is formed by losing another HCN group. The ions at 67 m/z peak are generated from 82m/z precursors by the elimination of methyl group CH_3. The final fragment ions at 55m/z are formed by losing an HCN group. The caffeine fragmentation pathway and its mass spectrum are shown in Figure 10: Fragmentation pattern of caffeine. During the fragmentation process, the CF molecule undergoes several changes by breaking its chemical bonds. The most important recorded CF peaks are 194, 165, 137, and 109 m/z *[75]*..

Since CF major metabolites (PX, TB, TP) have comparable chemical structures (i.e. they are isomers), they have the same molecular weight of 180 g/mol. When these compounds are detected and analyzed using mass spectroscopy-based techniques such as EI-MS and GC/EI-TOF-MS, they show several significant product ions in protonated mode $[M+H]^+$ at 180 m/z, 237 m/z and 125 m/z for paraxanthine (PX); for theobromine (TB) at 181 m/z with fragmentation pattern 137m/z, 109m/z, 67m/z and 55 m/z; and for theophylline (TP) at 180 m/z with fragmentation pattern 123 m/z, 95m/z and 86m/z [74]. Figure 11 Illustrates the mass spectra and their peaks for PX, TB and TP and respectively.

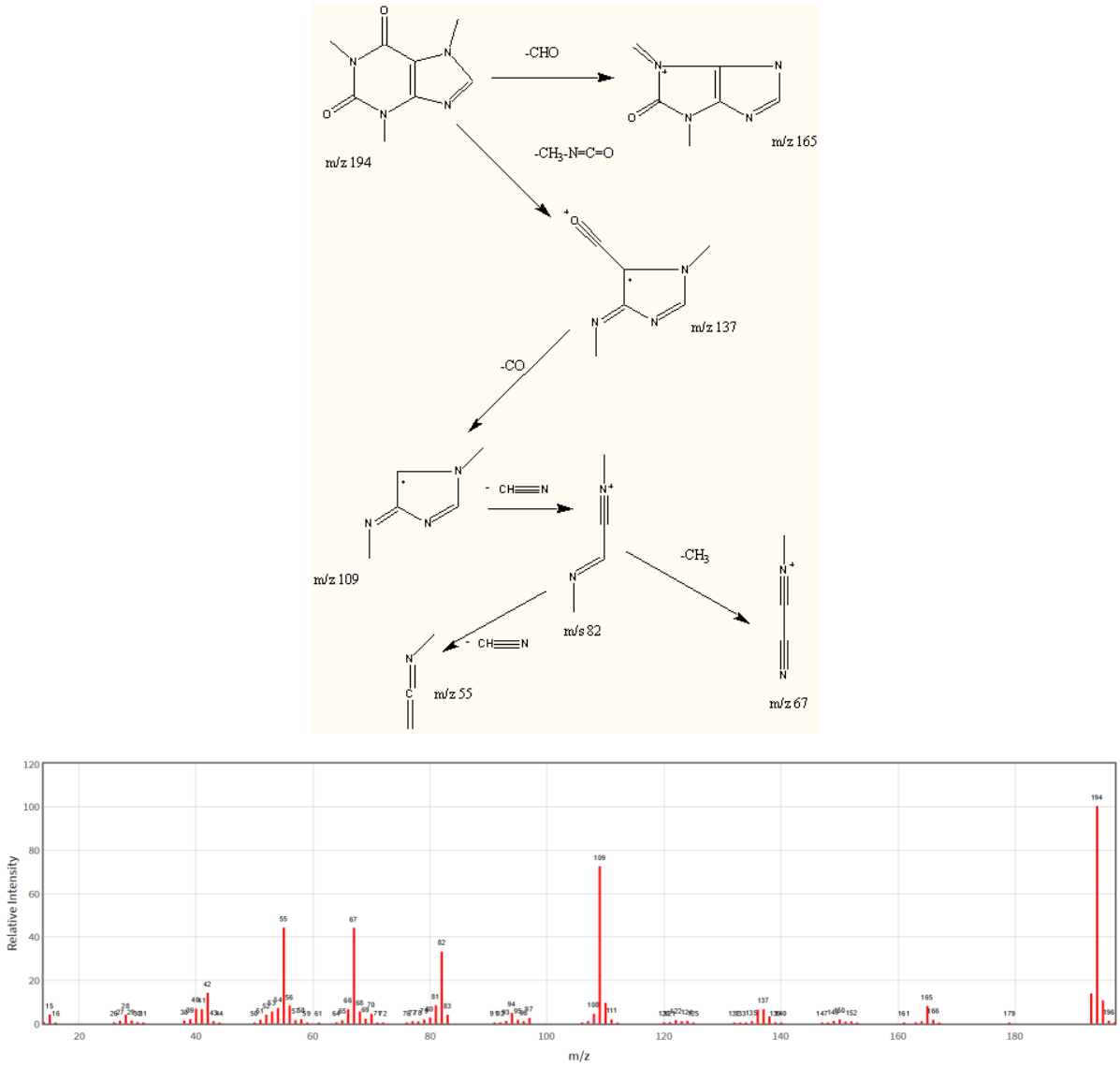


Figure 10: Fragmentation pattern of caffeine. During the fragmentation process, the CF molecule undergoes several changes by breaking its chemical bonds. The most important recorded CF peaks are 194, 165, 137, and 109 m/z [75].

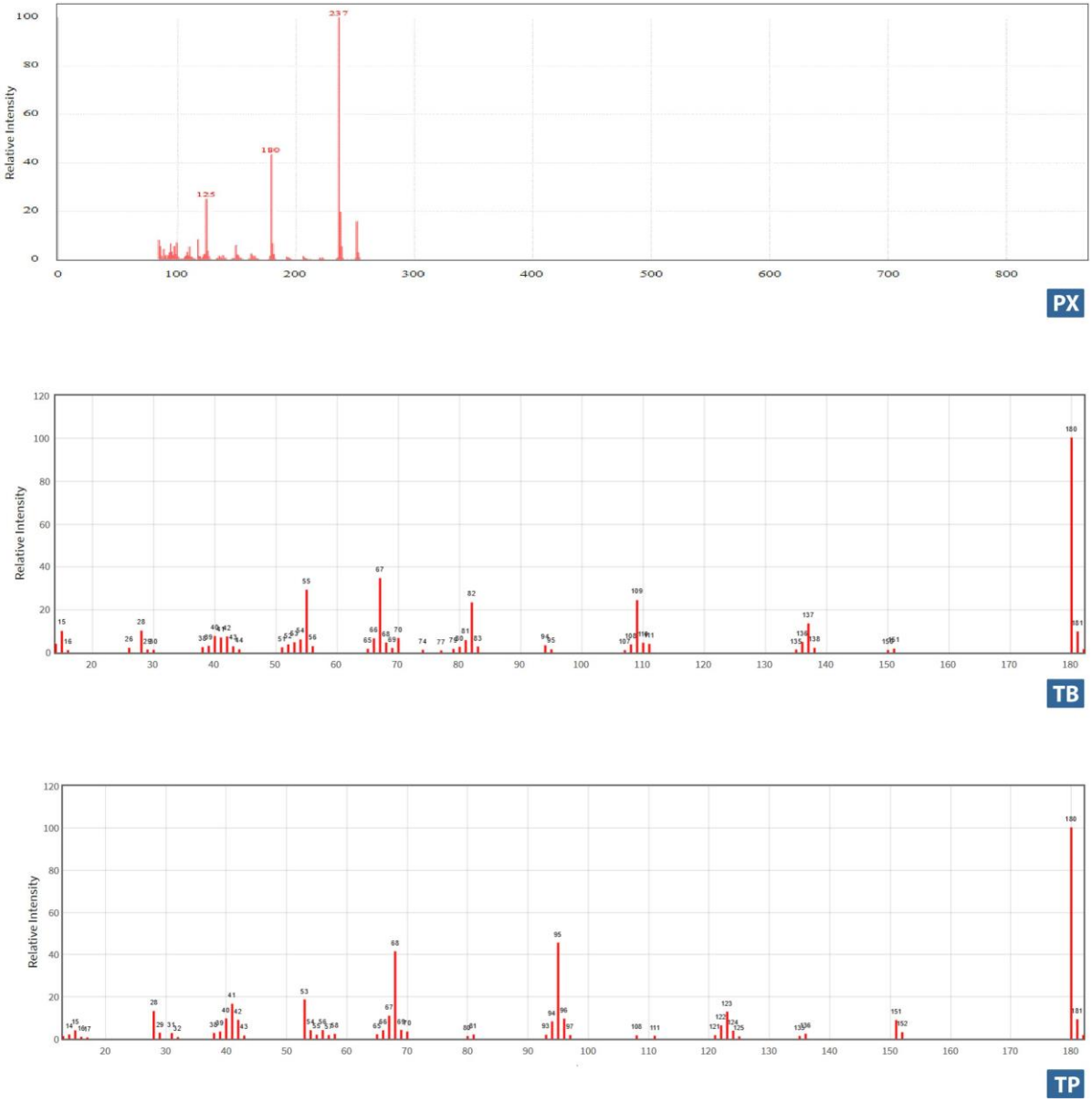


Figure 11: Shows mass spectra for caffeine three major metabolites (PX, TB, TP) in protonated mode $[M+H]^+$ with most important fragments values in m/z form and their intensities [74].

3 Statistics and hyperspectral imaging analysis

3.1 Introduction

Hyperspectral imaging (HSI), also called chemical imaging, is one of the spectral imaging technologies that has been used widely and successfully in environmental monitoring, resource assessment, and other remote sensing domains in order to combine the spatial and spectral information of the chosen area [76]. HSI integrates conventional imaging and spectroscopy to obtain three-dimensional data sets containing both spatial and spectral information of the sample. In the last few decades, mass spectrometry imaging based on the HSI method started being used in the statistical field [77], as well as the biological and pharmaceutical fields [78]. Many ionization technologies are successfully involved in mass spectrometry imaging applications. For example, Secondary Ion Mass Spectrometry imaging (SIMS imaging) was used to prove the stability of solid-state peptides, proteins, and biopolymers by detecting leuprorelin peptides drug distribution using a matrix of hydroxypropyl-cellulose [79]. Moreover, MALDI MS imaging has proven to be a valuable technique for proteomic analyzing and for imaging latent fingermarks by detecting trace materials within these prints and simultaneously visualizing their spatial distribution.

HSI was originally developed for remote sensing applications by using satellite imaging data of the earth, but nowadays is applied in diverse fields such as food science, pharmaceuticals, and medical diagnostics [80]. Hyperspectral imaging principles are analogous to a stack of images. Each one is acquired at a narrow spectral band, telling us the components' essence and where they are located. HSI can process and analyze information across the electromagnetic spectrum ranges, including ultraviolet (UV), visible (VIS), near infrared (NIR), mid infrared (IR), and thermal infrared ranges (TMS) [81].

In HSI, the acquired data of an image is represented in a hyperspectral cube that contains thousands of pixels. Each pixel represents the whole spectrum in an m/z versus intensity plot [82]. However, the generated dataset contains a huge amount of data which are distributed in a high-dimensional space. This makes the visual extraction of the desired information difficult. That's why statistical analysis algorithms are required [83]. To detect the CF molecules and lipids and their distribution in the fingerprint, HSI provides simultaneous observations of the spatial and spectral information of the analyte by accumulating temporal spectra of all single incidents for each image pixel [84].

The electromagnetic spectrum is shown in Figure 12. In these regions the reflectance, transmission, photoluminescence, or Raman scattering can be recorded by a hyperspectral camera with a spectral resolution similar to the miniature spectrographs.

Advantages of HSI:

- Reduction of human error
- Fast data acquisition
- No prior knowledge or the sample is required
- Selectivity can be achieved by means of multivariate statistics
- Ability to illustrate the results
- More detailed images

Disadvantages of HSI:

- Needs fast computers
- Sensitive detectors
- Large data storage capacities are needed for analyzing hyperspectral data since hyperspectral cubes are large, multidimensional datasets likely exceeding hundreds of megabytes

3.2 Fundamentals of hyperspectral imaging

The use of hyperspectral imaging began in the 1970s and 1980s for minerals mapping. During those years, HSI underwent a series of developments in terms of hardware, software, and computing power. Early usage of HSI combined with mass spectrometry has been conducted at NASA Jet Propulsion Laboratory (JPL) by Alexander Goetz and his coworkers using an airborne imaging spectrometer in the 1980s [85]. Nowadays, HSI application has been extended beyond remote sensing and control and are used in various scientific and research fields due to their accurate, nondestructive methods and ability to provide high-resolution images.

HSI combines spectroscopy and the power of digital imaging, allowing a greater chance to obtain clear and detailed images where every pixel contains thousands of contiguous spectral bands. The importance of HSI technique comes from its ability to determine the chemical composition of the sample since it offers more spectral information, which is useful for identification and quantification purposes [86].

In order to deal with hyperspectral imaging technique, it's necessary to describe the theory behind it. Thus, some of the essential expressions regarding spectroscopy will be discussed in this thesis such as the electromagnetic spectrum, the light behavior, and its properties.

According to quantum physics, the light has different properties and can be both waves and particles. When the light acts like a wave, its speed depends on wavelength and frequency, as well as some resulting behaviors including reflection, refraction, and diffraction. When light is described as particles, the energy it carries is related to the wave frequency given by Plank's relation:

$$E = h \gamma$$

Where E is the photon energy, h is Plank's constant (6.626×10^{-34} J.s), and γ is the frequency. This relation illustrates that light behavior is partially dependent on the amount of energy it carries. The spectroscopy field was developed to study light characteristics using instruments to analyze the light spectra. Thus, it played an essential role in discovering the molecules' properties during their interaction with light. Since hyperspectral imaging is applied in electromagnetic spectrum ranges including the visible range, more information and visualization can be extracted and recognized than the human eye.

3.2.1 Basics of Spectroscopy

The first concept of spectroscopic technique was invented in 1665 by Sir Isaac Newton when he simply passed light through a prism, splitting the light into multiple colors. He described the concept of light dispersion and the optomechanical hardware of a spectrometer [87]. In spectroscopy, physical characteristics such as reflectance, transmittance, and absorbance that result from the interaction between the electromagnetic radiation and the sample give us quantitative and qualitative information about the sample [88]. The basic principle shared by all spectroscopic techniques is to hit the sample of interest with a beam of electromagnetic radiation and then observe the sample's response to that stimulus as a function of the wavelength. Thus, the materials are recognized based on their different spectral signatures since each material has its unique spectrum.

An electromagnetic spectrum consists of a full range of frequencies with their respective wavelengths and photon energies. It consists of electric and magnetic field components that oscillate in phase perpendicular to each other and to the energy propagation direction. The frequencies of the electromagnetic waves cover a range up to 10^{25} hertz. In other words, the wavelength will range from thousands of kilometers long down to the size of the nucleus of an atom as shown in Figure 12. These electromagnetic waves are classified according to their frequency into different groups from low frequencies (long wavelengths) to high frequencies (short wavelengths) respectively: radio waves, microwaves, terahertz waves, infrared, visible light, ultraviolet, X- ray, and gamma rays [89]. In each group, the electromagnetic waves have different behaviors due to their different characteristics. However, these categories sometimes overlap. Therefore, different technological applications are used in various fields.

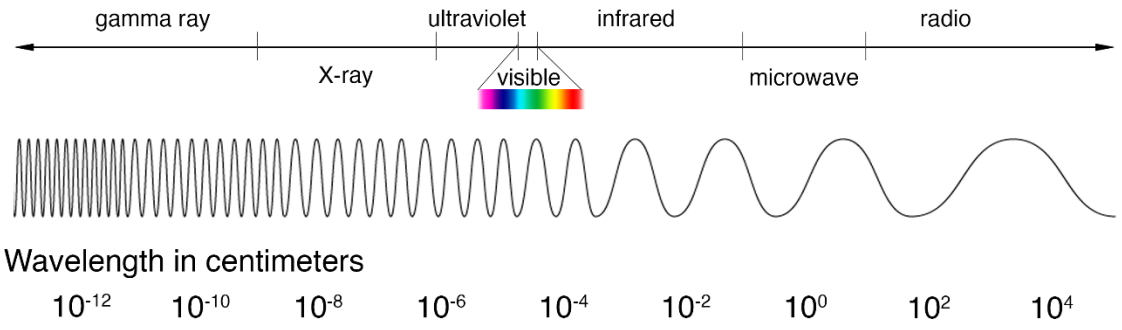


Figure 12: Parts of an electromagnetic spectrum that hyperspectral images can be obtained. In visible light, a 3D data cube is an RGB color image where each pixel has red, green and blue color, while the invisible hyperspectral image range can extend beyond the visible range (ultraviolet, infrared) [90].

Among the wide spectroscopic ranges, ultraviolet spectroscopy was the best choice in this project for caffeine and lipids detection since they have a high UV light absorption. This is done in MALDI MS by using UV pulsed laser that interacts with the sample, causing physical and chemical changes depending on the amount of energies they absorb and the selected wavelength.

3.2.1.1 Interaction of light with matter

Hyperspectral imaging technique uses the interaction of light with matter to determine the physical features and characteristics of the materials based on the optical properties of such an interaction. Figure 13 illustrates different types of interactions between light and matter, starting from the simplest, called reflection, which happens on the sample's surface and may give some information about the sample. Upon entering the sample, the light can be scattered or absorbed.

This scattering happens when the light changes its direction from the straight trajectory in which it enters the sample, causing light deviation by a specific angle called scattering angle. Scattering depends on two main factors: the light wavelength and the size of sample's particles. The electromagnetic scattering can be divided into two types: elastic scattering, which describes a process where the total kinetic energy of the particles doesn't change but their propagation direction is deviated; and inelastic scattering, where a small amount of incident particles scatter when the vibration state of the sample molecules cause shifting of the corresponding wavelength, causing some of the energy of the incident particles to be lost or increased. One important application of inelastic scattering of photons in a spectroscope is called Raman scattering, which can be used to chemically analyze the scattering sample by measuring the vibrational state of the molecules.

At last, absorption occurs when the total energy of the photon is taken up by the matter. However, this electromagnetic energy is converted to different types of internal energy of the absorber, a kind of property that is wavelength dependent. The absorption in the visible and ultraviolet spectra corresponds to the electronic transition in the molecules. While in near infrared and infrared ranges, the absorption depends on the vibrational modes of the molecules [91]. After excitation, the molecules de-excite and release energy in the form of radiation (such as heat or photoluminescence) or transfer this energy to other molecules.

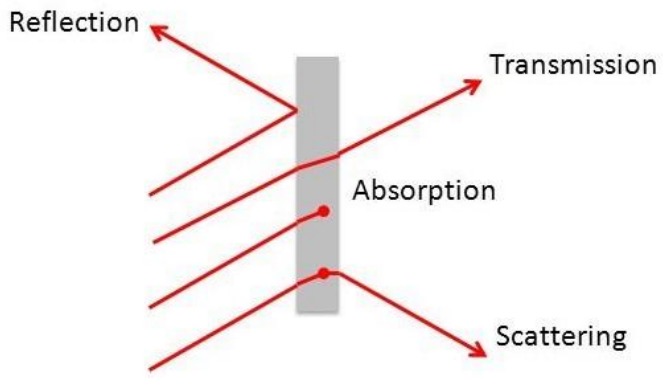


Figure 13: The interaction of light with the sample results in many physical phenomena: reflection, transmission, absorption, and scattering. These phenomena form the basic concept behind hyperspectral imaging principles.

3.2.1.2 Hyperspectral image acquisition

A hypercube is similar to a group of stacked images, each representing a narrow spectral band. The obtained dataset is a three-dimensional block of data. The first and second dimensions (x, y) are called spatial dimension and the third one is called the spectral dimension (λ). HSI offers researchers to ability to see beyond RGB image planes, meaning higher spectral resolution and more details of the heterogeneous samples. A hypercube formation is depicted in Figure 14.

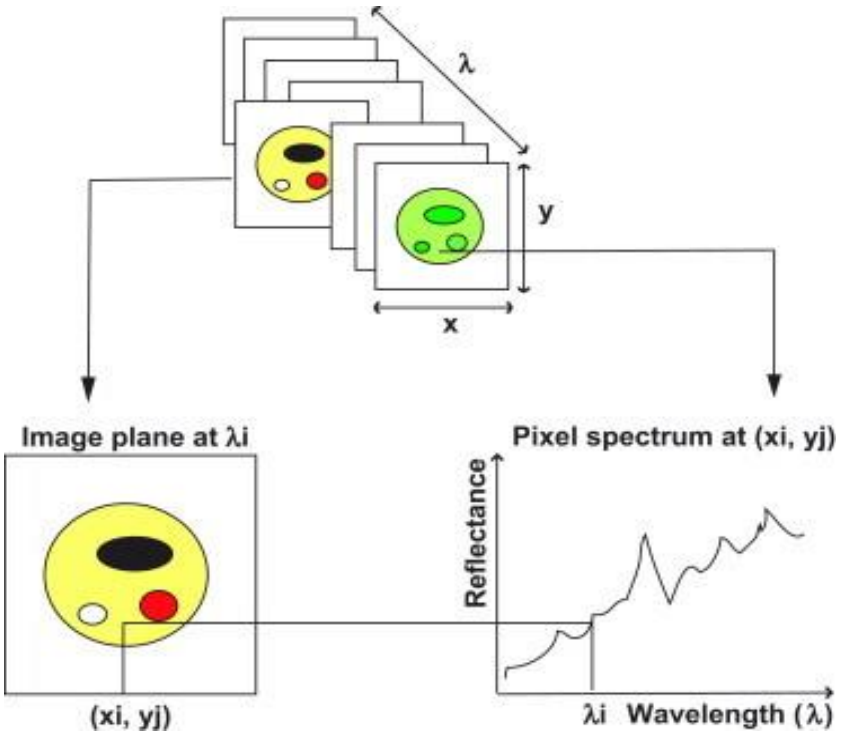


Figure 14: A hypercube is formed from stacking many images. Each image plane has two spatial dimensions (x,y) at a particular wavelength (λ) and each image pixel (x_i,y_j) represents the whole spectrum at a specific wavelength (λ_i) [93].

Therefore, each wavelength (λ) is corresponding to an image in the hypercube, and each pixel in this cube gives a corresponding spectrum. By using this concept, we end up with thousands (or more) of spectra, each one carrying spectral signatures of the sample. However, the spectral imaging depends on how many spectral bands are being used. If several spectral bands are used, for example, then multispectral imaging is obtained, while HSI uses hundreds of spectral bands.

There are three strategies for three-dimensional hypercube acquisition. All of them work through temporal scanning by accumulating two-dimensional data in sequences due to the inability of obtaining information in all three dimensions in one time. These ways of acquiring a hypercube are commonly known as point scanning (whiskbroom), line scanning (pushbroom), and area scanning (staredown).

Point scanning (whiskbroom) works on the full spectrum acquisition in every single point and requires either displacing the sample or moving the camera and keeping the sample in fixed position. The light of each point is subjected to a spectral analysis. Once this process is finished, the second spectrum of another point is recorded. This type of scanning is obtained in spatial directions (x and y) until the hypercube is completed. Point scanning is clarified in Figure 15 (a).

Line scanning (pushbroom) deals with the one image line including all spectra of all pixels simultaneously. The light is detected in a two-dimensional charge-coupled device (CCD) detector. This way, the data matrix with two dimensions (spectral and spatial) is acquired as shown in Figure 15 (b). The second spatial dimension is obtained by moving the detector across the sample's surface orthogonally to the imaging line. This is achieved by either moving the sample and keeping the hyperspectral camera fixed, or by moving the camera while keeping the sample in fixed position [94].

Area scanning (staredown) is when a complete hypercube is obtained by collecting a sequence of images with two spatial axes and one wavelength band (spectral axis) at a specific time. There is no need to move the sample. A tunable filter is used to modulate the wavelength of the incoming light. Figure 15 (c) shows area scanning modes.

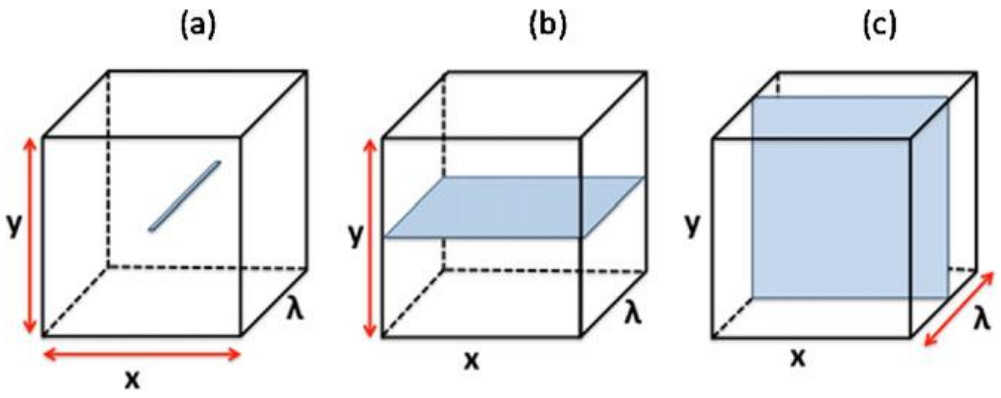


Figure 15: Acquisition algorithms for a three-dimensional hypercube: (a) point scanning, (b) line scanning, and (c) area scanning [86].

3.3 Preprocessing of hyperspectral images

As explained in the previous section, the hyperspectral data cube consists of thousands or millions of data points. Each pixel is highly correlated to its neighboring pixels. The acquired image may contain many noises and suspicious pixels including the non-informative background. There are several factors which cause such undesired pixels, including the instrument itself such as during the scanning step where the detector may generate some suspicious pixels. Moreover, the behavior of radiation methods in spectroscopy may create some artifacts. Thus, all these factors make the preprocessing of the datasets before analyzing them an essential issue. Multivariate analysis techniques such as principle component analysis (PCA) show high performance regarding the extraction of the useful data.

In a three-dimensional HSI cube the spatial dimensions (x,y) consist of columns and rows coordinates, and the spectral dimension consist of N wavelength which can give the spectral signature of chemical

component of the complex sample. Each pixel (x_i,y_j) of the three-dimensional cube represents a spectrum over thousands of layers (m/z values) with their given intensities. Each pixel in HSI consists of more than 20 bands. The image is a gray scale image stored in the computer as 8-bit integers, giving a range of possible values from 0 for black and 255 for white. Based on the experiment, several gray levels are used such as 12 bits, 14 bits, and 16 bits.

In order to process our data, the hypercube is transformed into a two-dimensional array of spectral vectors by using many algorithms of multivariate data analysis. Converting the hypercube to a two-dimensional matrix is shown in Figure 16.

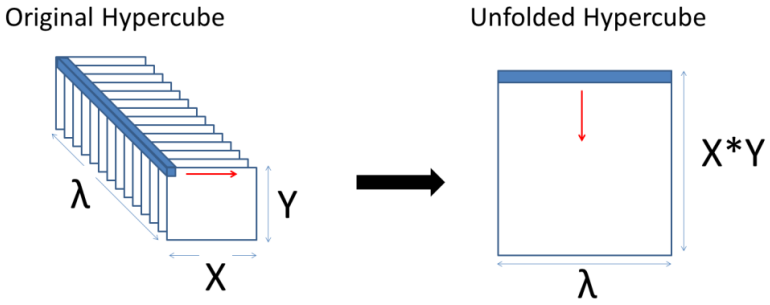


Figure 16: Unfolding the 3D cube to a 2D matrix where each pixel (x_i,y_i) represents a whole spectrum and this spectrum contains the m/z values with their intensities [105].

In order to process our obtained hyperspectral data from the instrument, special software has to be chosen. ImageLab[1] *(version 2.91)* software is used since it offers support for many spectroscopic imaging techniques in different domains, most importantly in mass spectrometry. Built-in statistical methods such as PCA, multiple linear regression, and several statistical tests can be used to analyze hyperspectral images. To apply the statistical tests, data visualization, and preprocessing, DataLab [2] *(version 3.530)* software is used.

3.4 HSI analysis based on machine learning algorithms

Since hyperspectral images contain high dimensional informative data as well as dead pixels (background pixels) and noise that come from many factors during the experiment, it becomes meaningless to analyze the data manually. Thus, many machine learning procedures are used to

[1] http://www.imagelab.at/
[2] http://datalab.epina.at/

automatically analyze the huge amount of data which are distributed in thousands of dimensions. Generally, machine learning algorithms are divided into two categories: unsupervised and supervised learning. These two classes have some differences.

In supervised learning, the machine should learn from datasets, called training datasets, to make correct predictions. The known training dataset T involves different x variables, called input or independent variables, and y values (labeled variables), called dependent variables. Specific function ƒ in the training dataset maps an input to an output based on input-output pairs, and then the inferred function is used to map new pairs. Moreover, this function is affected by training set size (i.e. the larger the training set, the more learnable the function ƒ).

Supervised learning can be grouped into two further categories: regression analysis allows to observe the relation between the dependent variables and the predictors, the dependent variables have to be continuous. While in classification, the output variables are discrete, and it works on identifying the class that the data belongs to.

In unsupervised learning, there is no direct supervision by human, so no labeled y variables are needed. Thus, the system classifies the data based on similarities or differences between them without any prior training. The most common unsupervised learning methods are cluster analysis and principal component analysis (PCA).

In this work, we will try to apply some algorithms in order to discriminate between two groups: “non-caffeine and caffeine” groups based on the caffeine which shows up in the fingerprints, and “country of origin” of two groups of individuals based on the variation of the lipids in their fingerprints.

3.5 Curse of dimensionality

Dealing with HSI is a problematic task due to the distribution of the information in high dimensional space, making the data predictions, classification, and decision boundaries hard to achieve. This is called the curse of dimensionality.

PCA applications vary in several areas including data compression, image processing, visualization, exploratory data analysis, pattern recognition, fingerprint-based access control, and time series prediction [95]. This can be used in our work to differentiate between groups, since we can use variance to get the general difference instead of using pre-determined features to feed clustering/classification algorithms.

Having large numbers of variables can present some problems by possibly overfitting the model to the data or violating assumptions of some research hypothesis. Thus, it’s necessary to understand the relationships between those variables and identify which ones are most important. This is usually achieved by reducing the number of variables and selecting the most influenced ones [96]. Technically, this is called “dimensionality reduction.”

PCA is one of the more powerful ways to reduce dimensions. It is a multivariate technique to analyze data in which observations can be described using correlated variables. Using linear transformation, PCA converts the observations set of possibly correlated variables into a set of values described by linearly uncorrelated variables, called principal components. PCA is one of the most popular multivariate statistical techniques.

The goals of PCA are to:

- Extract the most important information from the data table
- Compress the size of the data set by keeping only this important information
- Simplify the description of the data set.
- Analyze the structure of the observations and the variables

PCA procedure begins with calculating data means and subtracting them from the data, then calculating covariance, correlation, or scattering matrix C, which represents the statistical relationship between variables. After that, eigenvalue decomposition of C matrix is computed. This results in the following relation:

$$C.E = E.\Lambda$$

$$E = (e_1, \dots., e_p)$$

$$\Lambda = (\lambda_1, \dots., \lambda_p)$$

If we retain only k eigenvectors which correspond to the highest eigenvalues, then we can reduce problem dimensionality.

Eigenvectors represent the projections (direction) which we project the original vector along, while eigenvalues represent strength or scale (variance of amount of information) a specific projection can offer if the data is projected in that specific direction. The higher the eigenvalue, the more variance in that specific direction. Therefore, the amount of overall information is directly related to cumulative contribution of eigenvalues.

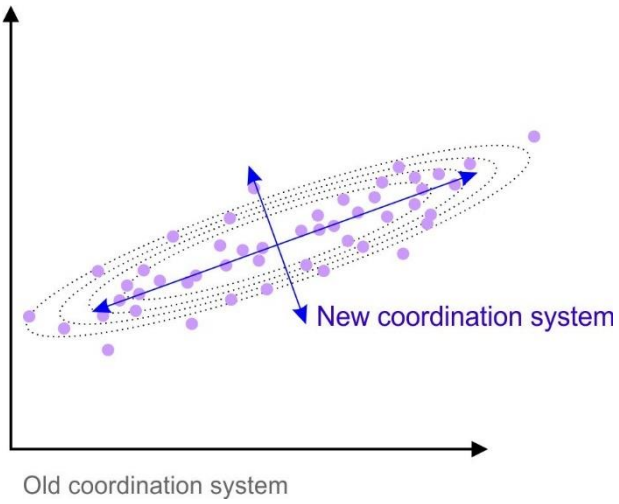


Figure 17: Shows principal component analysis coordination conversion, where the new coordination contains the highest variance.

It is always recommended to use a correlation matrix when we are dealing with different variables scales so we can avoid the influence of large scale variables on the whole dataset.

In our case, we can notice high intensities around 158 m/z which corresponds to the using matrix in our experiment called 1,5-diaminonaphthalene (DAN) matrix. The variables in this region will contribute highly to the variance of the whole dataset and the direction of the new principal components, but this is not correct because these applied features are not related to the problem. Therefore, standardization scaling is applied using a correlation matrix.

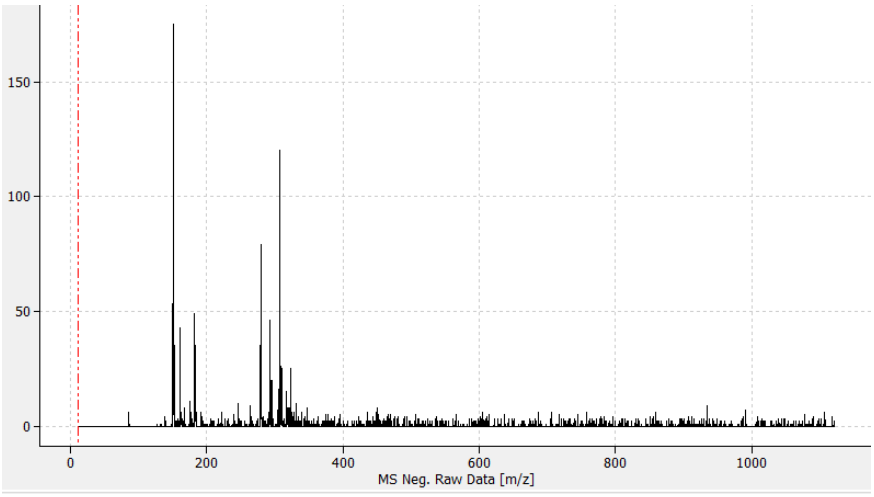


Figure 18: A fingerprint spectrum shows the high intensity of the used matrix with values around 158 m/z.

3.6 Variable selection

One of the largest difficulties in HSI processing is the random distribution of the information in thousands of dimensions and the existence of many redundant or irrelevant features, making it hard to analyze the data. Thus, applying the variable selection task is essential to achieve a correct prediction. Variable selection concept works by selecting a subset from the original data set and removing all the redundant features without losing much information. The importance of variable selection comes from many factors: it overcomes the curse of dimensionality problem, makes the model easier to handle, shortens processing time, and provides easier interpretation, and most importantly, increases model stability and robustness.

Generally, the variable selection can be grouped into three main algorithms:

- Wrapper methods use the subset evaluator to create all the possible subsets from the original data. They then apply the classification algorithm for each subset and choose the subset where the classification algorithm performs the best. Stepwise regression is the most common type of wrapper, which is used broadly in statistics and will be used in this work for applying multilinear regression method.
- Filter methods usually work by selecting the best subset independently from any learning algorithm. Thus, it is less time consuming and has lower prediction performance than the wrappers.

- Embedded methods involve choosing the best subset here is depending on the classifier which have specific selection methods.

Stepwise selection is a combination of the forward and backward selection methods. First, it adds the variables and then checks their significance according to predefined level. If they are non-significant it will remove them. This creates a robust classifier and consequently a correct prediction.

3.7 Multiple linear regression

In essence, multi linear regression is basically the same as linear regression but applied on higher dimensions in space. Linear Regression is the determination of an equation coefficients which can relate response measurements set Y to descriptors set X:

$$Y(x) = \alpha.X + \beta$$

Where α is the slope or the "coefficient", and β is the intercept or the offset of the vector.

This concept can be generalized to any number of features p, and is called multi linear regression:

$$Y(x) = \alpha_1.x_1 + \alpha_2.x_2 + \alpha_3.x_3 \dots\dots\dots + \alpha_p.x_p + \beta$$

$$Y(X_p) = \sum^{p} (\alpha_i.x_i) + \beta$$

The fitting error term e_i can also be added to each observation to achieve an exact value:

$$Y_i(x) = \alpha_1.x_1 + \alpha_2.x_2 + \alpha_3.x_3 \dots\dots\dots + \alpha_p.x_p + \beta + e_i$$

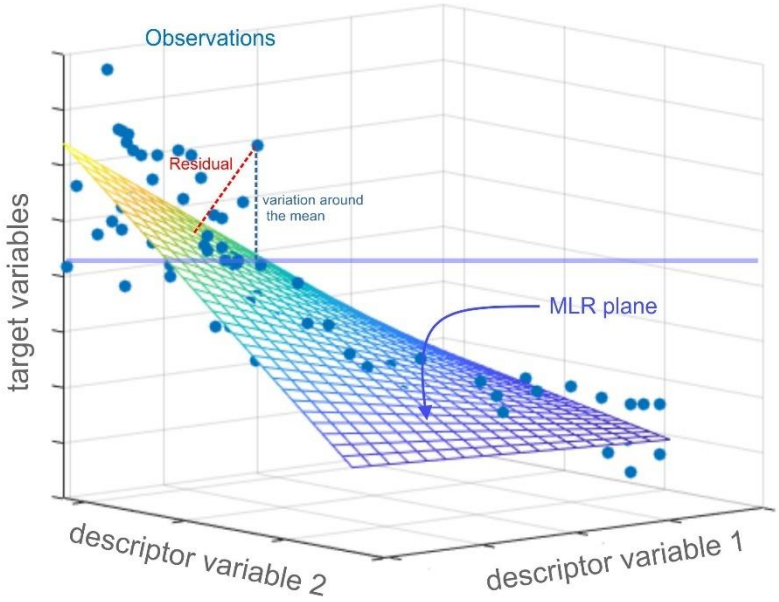


Figure 19: Shows multi linear regression plane, observations, and residual of one observation and distance from the mean.

The selection of best fit (p-plane) of the data is described by three main terms: R^2, F-value, and P-value.

R^2 can be seen as the quotient of the variation in the dependent target feature explained by the independent descriptor variables and the variation in the dependent target feature. Without taking descriptors into account, a larger value will result in the best p-plane.

F-values describe the quotient of the variation in the dependent target feature explained by the independent descriptor variables and the variation in the dependent target feature not explained by descriptors. F-value (and corresponding p-value) determines how the relationship is reliable, while p-value must be as small as possible.

Assuming we have P feature and N number of O observation then:

$$var(mean) = \frac{sum\ of\ squared\ residuals(Y_n)}{n} = \frac{\sum^N (y_n - \mu)^2}{n}$$

$$var(mean) = \frac{SS(mean)}{n}$$

$$var(fit) = \frac{sum\ of\ squared\ around\ least\ squared\ fit(Y_n)}{n} = \frac{\sum^N (o_n - \sum^p (\alpha_i . x_i) + \beta\ )^2}{n}$$

$$var(fit) = \frac{SS(fit)}{n}$$

$$R^2 = \frac{var(mean) - var(fit)}{var(mean)} = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

$$F = \frac{SS(mean) - SS(fit)/(p_{fit} - p_{mean})}{SS(fit)/(n - p_{fit})}$$

3.8 Partial Least Squares discriminant analysis (PLS/DA)

Among several multivariate classification algorithms, Partial Least Squares discriminant analysis creates a model to discriminate between Dependent variables and one or more target variables.

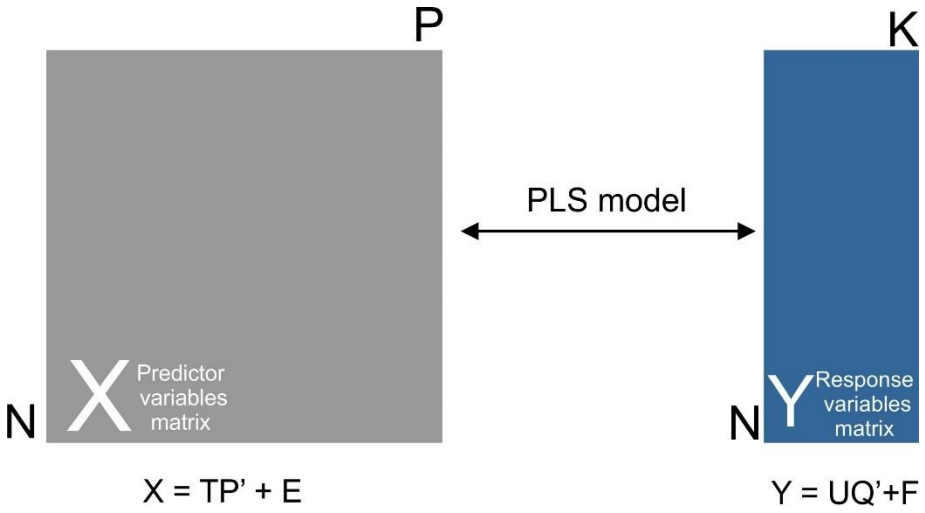


Figure 20: Partial Least Squares discriminant analysis links between multivariate predictor matrix and a multivariate response matrix.

Assume P predictor variables matrix X and K response variables matrix Y, with N number of observations. PLS calculates the scores and the loading of both X and Y in such a way that the first score in X matrix t_1 has maximum covariance r_1 with the first score in Y matrix u_1, meaning we can predict the first score in Y from the first score in X. This can be written as:

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ \vdots \\ r_n \end{bmatrix} . \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ \vdots \\ t_n \end{bmatrix}$$

Although this concept sounds similar to PCA, PCA searches for the maximum variance in both X and Y. In PLS, we are looking for directions X and Y in which we have best correlation between X scores and Y scores.

It is important to determine how many components we should include in the model. This can be achieved by examining different measurement error criteria such as the root mean square of residuals, and then choosing a model with the number of components that correspond to least error.

3.9 Overfitting

Overfitting problem can be described as fitting most or all of the data exactly, or very closely, to observations. This often results in failure to generate a stable model which can correctly predict the wider dataset. This problem can also be seen as fitting the noise or residuals instead of searching for a reliable relationship between data points. In this case, the model starts to memorize the data set instead of learning.

Overfitting extent depends on the number of parameters in the model and the number of observations. After reaching the optimal parameters number that offers least error, the addition of new parameters can result in increasing the error instead of reducing it, such as when we apply the model to other datasets (like in case of cross validation) as shown in Figure 21. Overfitting problem extent also depends on model structure and complexity.

The rationale behind overfitting originates from, among other reasons, a possible correlation between parameters, dependence of the parameters on a certain timeframe, and even the augmentation of unneeded or noisy data.

Thus, checking model performance with the large dataset is essential to determine the optimal number of parameters to provide “valid model”.

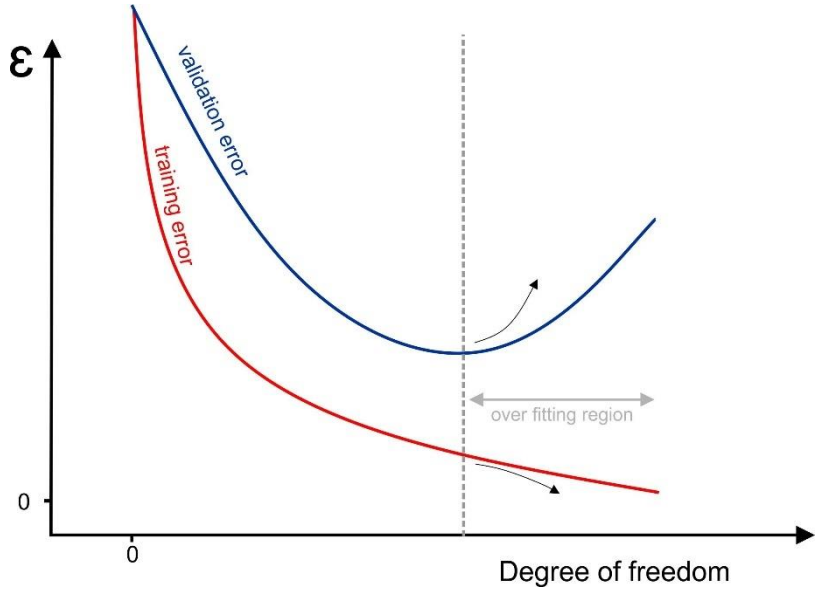


Figure 21: Shows overfitting problem region after a certain number of parameters, where the models will accumulate useful parameters to the model.

3.10 Cross validation

Cross validation is a procedure in which part of the whole dataset is used to estimate model parameters (train the model) while the rest is used to evaluate the model performance (test the model). This process checks the performance of the model on datasets different from the data used to generate the model, which helps to detect overfitting, biased observations models, and provides stable valid models.

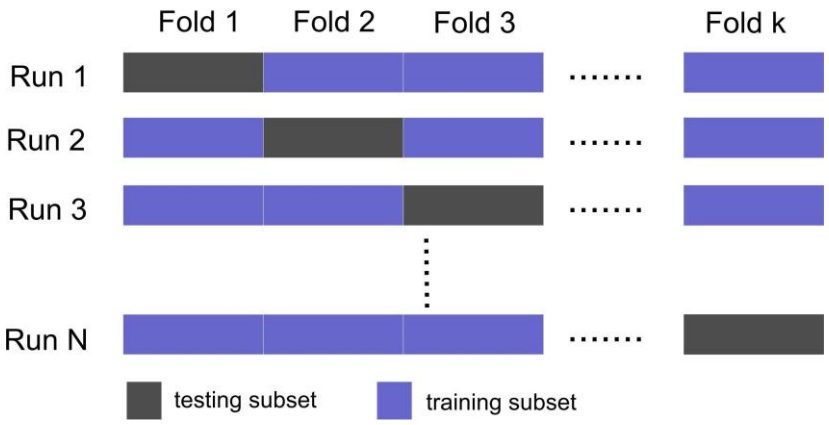


Figure 22: Cross validation concept where the data are divided into two subsets for training and testing in each run.

K-fold cross validation is a method in which we split the dataset into k-subset, estimate the coefficients using (k-1) subset, and then use subset k to evaluate the model. This process can be repeated N times to use all possible combination of training and test subsets, adapting the run with least evaluation error. This method can be used multiple times with a different number of parameters or components to determine the optimal ones that can be used without overfitting or over-parameterizing the models.

4 Methods

4.1 Introduction

MALDI MS is an acronym for Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry, it was invented and developed in the 1980s by two scientists from Germany named Michael Karas and Franz Hillenkamp [97]. MALDI is a soft ionization method that can analyze a wide range of molecules with high sensitivity and minimal fragmentation. This minimal fragmentation is achieved by mixing a material called the matrix with the sample, where the number of matrix molecules should be larger than number of analyte molecules

[98]. MALDI coupled to Time of Flight mass spectrometry (MALDI TOF MS) can be used to analyze and detect different types of large biomolecules like proteins, DNA, and peptides, along with other smaller molecules like lipids and fatty acids [99].

Soft ionization methods are favorable in biological-related analysis since they preserve the sample from destruction in contrast to other analysis methods that use hard ionization methods, destroying the sample. For example, MALDI MS can analyze the proteins after isolating them by gel electrophoresis or other isolation methods, and for oligonucleotides (DNA) synthesis studies [100] . Furthermore, it allows many immunological and biochemical experiments to identify microorganisms such as bacteria or fungi [101][102], antibiotic susceptibility [103], and even some cancerous regions in living tissues like pancreatitis cancer in microbiological laboratories [104]. These various applications can present an advantage for MALDI MS imaging-based analysis when it comes to working with biological substances like lipids, nitrogenous organic compounds like caffeine, and other human body metabolites.

4.2 MALDI TOF MS work principles

MALDI-TOF-MS methodology is a multi-step process:

1. **Ionization step**

First, the ionization process begins by mixing the analyte of interest with material called the matrix in specific amounts and under special conditions. Additional details are presented later on in this work. The mixture of matrix and analyte is irradiated using a pulsed laser beam. The number of matrix molecules should be larger than the analyte molecules to absorb the majority of the photon energy and protect the analyte molecules from clustering and fragmentation. This happens by forming matrix isolation [105]. Because MALDI technique works in vacuum surrounding, the matrix and analyte molecules are ablated from the plate and transformed to a gaseous phase, forming a plume of ions. The plume contains neutral and ionized matrix molecules that are clustered around the analyte molecules, ionizing them by transferring one or more protons from the matrix molecules to the analyte of interest. Adding or removing one or more protons is called protonation and deprotonation modes ($[M+H]^+$, $[M-H]^-$ respectively) The formation of ions is highly dependent on the ionization method being used, such as thermal ionization and gas-phase photoionization. A high voltage electric field is applied in this step to accelerate the molecular ions to the mass analyzer.

The mass spectrum is generally affected by several parameters during the ionization process such as the ablation process, the type of used matrix, the sample preparation methods, and the matrix-analyte combinations. All these parameters make the ionization step complicated to fully comprehend.

Choosing the matrix is important for obtaining the acquired soft ionization. This matrix should have a notable absorption coefficient at the applied wavelength. The optimal MALDI matrices are small organic molecules used to facilitate the ionization process. Usually, the matrix is a conjugated system like functional attached benzene rings, which is exposed to a pulsed UV laser with a wavelength of 337nm for 0.5 to 20 ns.

TOF mass analyzer

After particles ionization, the resulting matrix and analyte ions enter a draft region in the Time of Flight (TOF) mass analyzer where they are separated according to their velocity. This velocity is highly dependent on the ions' mass to charge ratio (i.e. heavier ions with the same charge will have lower velocity in contrast to lighter ions). Thus, the mass to charge (m/z) ratio is determined by the time it takes for the ions to reach the detector. This process is done under high vacuum to prevent collisions between ions. In this step, the ions are subjected to an electric field to impart a constant amount of kinetic energy to each ion. The smaller ions travel faster than the larger ones and can be recorded by the detector as shown in Figure 23

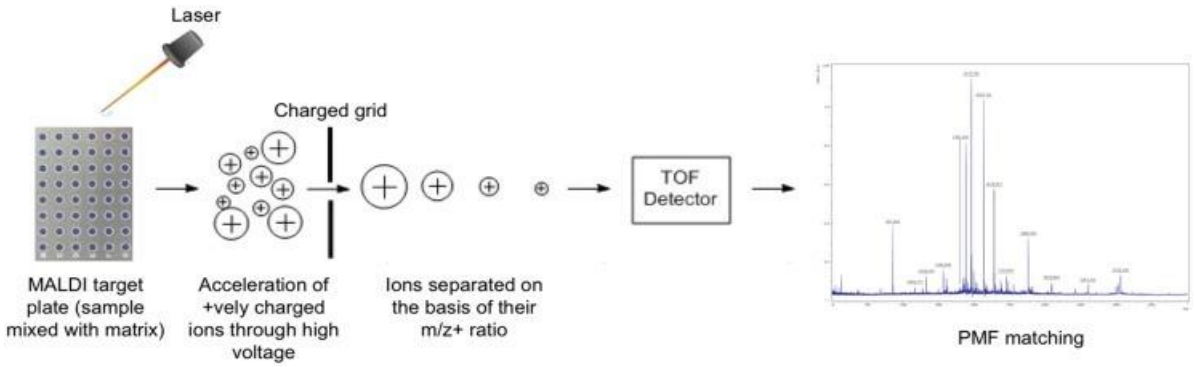


Figure 23: Schematic diagram showing the work-flow in a MALDI-TOF MS starting from the ionization chamber and then creating the ions. Then they enter the mass analyzer for ions separation based on their m/z ratio. Finally, these ions can be detected and a mass spectrum is formed [106].

2. Ion detection

MALDI MS instruments can be coupled with several detection methods, but two fundamental methods are mostly involved: the linear and reflection modes [107]. In the linear mode, the ions with low masses will arrive at the detector faster than the larger ones. But from a practical point of view, the molecular ions that have the same m/z don't receive the exact amount of kinetic energy;

they reach the detector at different times, affecting the spectrum resolution negatively. Nowadays, this problem has been overcome by using the reflection mode consisting of an electrostatic mirror that reflects the ions and sends them back through the flight tube until they reach the detector. The detector is positioned on the ion source side, opposite the mirror. The kinetic energy of the ions with same m/z will be corrected and will reach the detector at the same time. The reflection mode is shown in Figure 24.

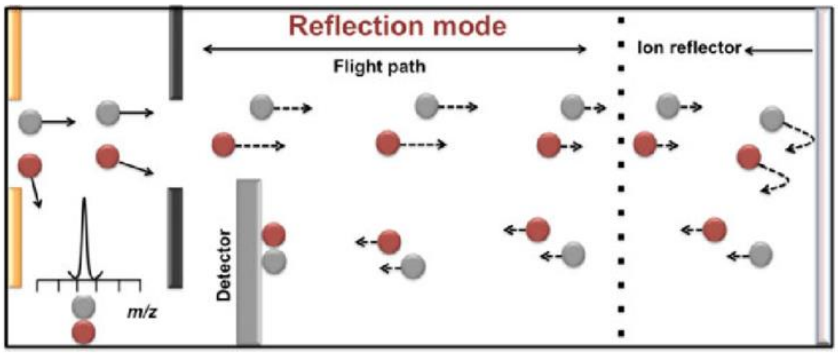


Figure 24: The reflection mode in Matrix-Assisted Laser Desorption/Ionization time of flight Mass Spectrometry [108].

Advantages of MALDI TOF MS:

- Fast and accurate identification of the molecules [109]
-
- MALDI saves processing time compared to the traditional techniques [110]
- The ability to ionize nonvolatile, large (MW> 200.000 Da) or polar molecules [111]
- The main property of MALDI is the soft ionization, meaning that the analytes of interest stay intact when they are ionized

Limitations of MALDI TOF MS:

- High costs for buying a MALDI-TOF MS instrument
- In MALDI MS the identification is usually limited by database knowledge, but this might be overridden by using a comprehensive library of spectra [112]
- Time-consuming for sample preparation to obtain good results

4.2.1 Laser in MALDI TOF MS

Before the invention of MALDI, some matrix-free approaches like Laser Desorption Ionization Mass Spectrometry (LDI MS) used infrared pulsed lasers to analyze unstable and nonvolatile biomolecules [113]. Today, after proving the capability of MALDI for wide classes of molecules detection, it is commonly applied to the infrared and ultraviolet ranges. Many studies have reported the efficiency of IR-MALDI as a soft ionization method and highlight its ability to minimize the degree of metastable ion fragmentation. They also refer to some limitations in analytical performance and requirements [114]. This limitation led to the development of an ultraviolet (UV) nitrogen pulsed laser beam with wavelengths close to the matrix's maximum UV absorption (around 337 nm) for commercial and research purposes. Because most of the MALDI matrices are optimized for UV wavelengths, the UV MALDI became widespread in many fields. It achieves good isolation and protection of the targeted analytes and is cheaper than other lasers including Er:YAG at 2.94 µm in the IR range [115].

One of the most important considerations in MALDI MSI is selecting the probation of the pulsed laser according to the type of the tissue or the analyte of interest (i.e. laser power, laser wavelength, and

laser spot size). In order to simultaneously achieve high signal intensities and avoid the signal overtones, the laser wavelength and the maximum absorption of the used matrix have to be matched [116]. The spot size has a notable influence on the desorption/ionization process and creates an amount of resulting ions [117]. The UV laser pulse duration is an additional factor here since it has a minimum effect on the ion intensities, usually within 0.3 to 20 ns and 266 to 355 nm wavelength [118]. To discriminate between the analyte signal and the matrix signal (blank value) which is called Limit of Detection (LOD) and defined as 3*standard deviation of the blank, a small square on the slide containing only matrix substance is predefined.

4.2.2 Matrix selection in MALDI TOF MS

As a part of MALDI technique, choosing the matrix is considered the cornerstone of MALDI principles. Many metabolites have been analyzed successfully based on finding a suitable matrix. We can apply the appropriate matrix to detect small molecules (MW<1.000Da) as well as macromolecules such as proteins and peptides. The matrix consists of molecules in crystal form mixed with cationization agents such as sodium and lithium, which encourage soft ionization and achieve a homogenous co-crystallization of the analyte of interest. Generally, two types of matrices are available: a wet matrix used for protein and peptides detection, and a dry matrix used for lipids. All matrices are organic, which has a significant influence on molecules analysis, specifically in negative modes such as N-(1-naphthyl), ethylenediamine dinitrate [119], 1,8-bis (dimethylamine) naphthalene (DMAN) [120], and 9-amino acridine (9-AA) [121].)[122]. Through the experiments, some of these matrices faced limitations in stabilization. This induced some scientists to find alternative substances. They started to use 1,5-diaminonaphthalene (DAN) matrix due to its radical hydrogen transfer capacity [123] and its ability to obtain high resolution images, especially for lipid imaging (the spatial resolution for lipids up to 10 µm). DAN is also effective for both positive and negative ion modes and is considered a main proton source to encourage ionization of the analytes. Based on these properties, DAN matrix shows good results for detecting and analyzing large molecules like proteins and peptides in positive mode and oligonucleotides in negative mode [124]. It has a molecular weight of 157.8 mg/mol, which is relatively low to allow easy vaporization. It is acidic, looks like colorless to pale purple crystals or lavender powder, and has the chemical formula $C_{10}H_{10}N_2$ [125].

Figure 25 shows a comparison between DAN and other matrices like α-cyanohydroxycinnamic acid (CHCA), 2,5-dihydroxybenzoic acid (DHB), and 9-aminoacridine (9-AA). DAN shows high performance for small molecules and metabolites with low molecular weight (MW<400 Da) like: M, malic acid (m/z133.014); G, glutamic acid (m/z 146.046); P, phosphoenolpyruvic acid (m/z 166.975); A, ascorbic acid (m/z 175.025); and U, UDP-glucose (m/z 565.047) in comparison to other matrices [126].

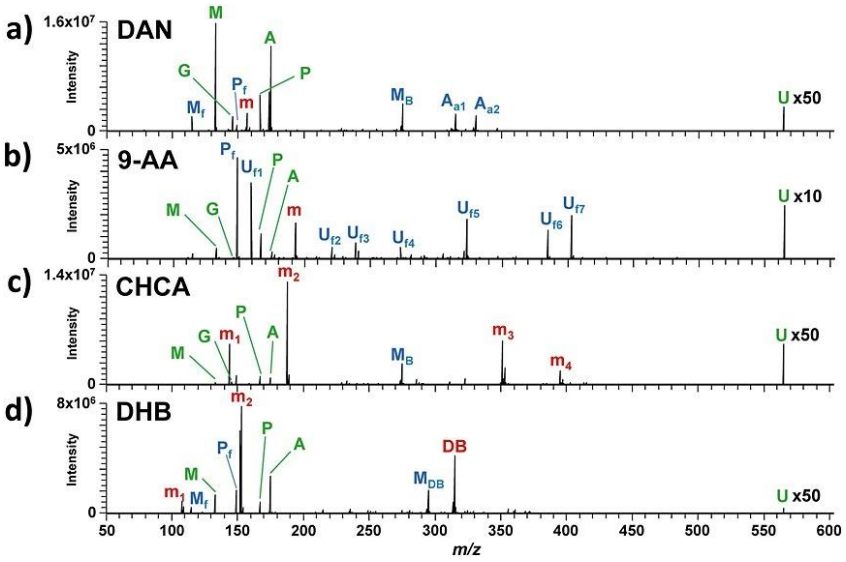


Figure 25: MALDI-MS spectra of a metabolites' standard mixture using several different matrices: DAN, 9-AA, CHCA, DHB. [135].

5 Experiments

In this chapter, the setup of the experiment will be described. Figure 26 shows the work flow for the experiment procedures starting from sample preparation, then introduces the fingerprint samples to MALDI MS for analyzing. After fingerprint data acquisition, some preprocessing procedures are applied to enhance the spectral data sets. Finally, some multivariate statistical models are applied to discriminate between caffeine and non-caffeine consumption and between the individuals county of origin based on the lipids variation on their fingerprint.

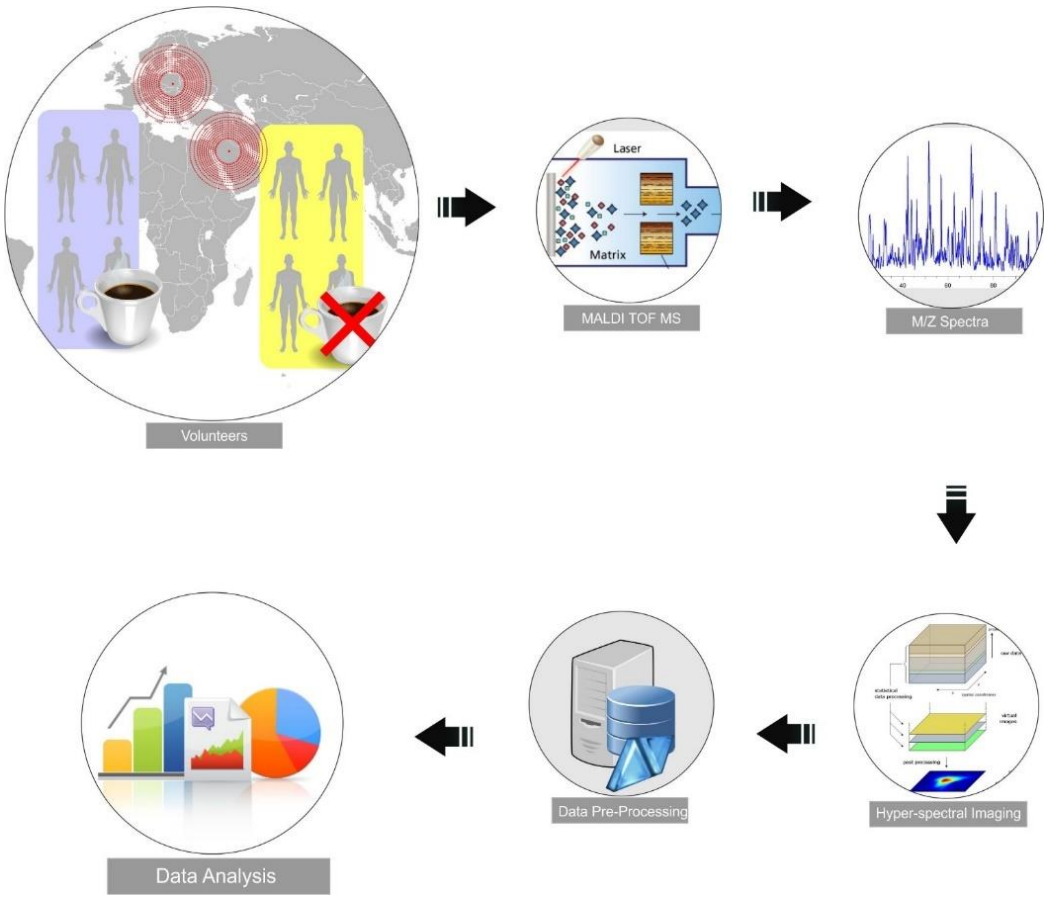


Figure 26: Experiment work flow for fingerprint data acquisition.

5.1 Sample preparation

In order to compare caffeine consumption and observe the significant information of the individual's region of origin based on the lipids and fatty acids in their fingerprints, two groups were created: "caffeine and non-caffeine" and "Austrian and Syrian" groups. For this purpose, six volunteers (three Austrian and three Syrian, all male and ages between 25 and 30) were asked to donate their fingermarks. The same conditions were applied to the six volunteers in order to get comparable information (i.e. keeping the general parameters as constant as possible for all individuals). To distinguish the chemical features from CF in the volunteers' fingerprints and prevent overlapping from other caffeine sources, they were asked not to consume any caffeine products from any source at least six hours before the experiment. They wake up and washed their hands with water only, without using any soap products since the detergents form spherical micellar structures, interacting with lipids

and eliminating them [127]. They were also asked not to use any type of cream because it causes the fingerprint to become smeared and contaminated, making it difficult to achieve the clear image. After washing hands, the finger is then wiped with isopropanol solvent to remove any contaminates that may be found from the ambient atmosphere.

After these instructions, the volunteers were asked to donate the first fingerprint without caffeine consumption on the Indium Tin Oxide coated glass slide (ITO). ITO is used due to its major characteristics of electrical conductivity and optical transparency [128]. After the first fingerprint acquisition, we asked the volunteers to drink two cups of coffee, equivalent to 0.25L without any milk or sugar (Kazaar Nespresso coffee blend has been chosen since it contains of approximately 125 mg of caffeine per capsule which is one of the most intensive choices)[3]. Volunteers waited about three to five hours after coffee intake, the average time for caffeine metabolism without touching anything to keep all the lipids and fatty acid and CF metabolites on the finger surface. Finally, after a set time limit, they donated the second fingerprint with CF in a new ITO coated glass slide. The total sample numbers are 12 fingermarks.

5.1.1 The sublimation process of DAN matrix in MALDI MS

In order to detect the desired chemical analytes with their unique spatial distribution in the deposited fingerprints, a dry and homogeneous matrix deposition is required to obtain a soft ionization with a maximum of UV light absorption. To achieve that, the sublimation protocol is a fundamental step to ensure the matrix deposition on the fingermark's slides. An amount of 25.5 mg of DAN (1,5-Diaminonaphthalene) is weighed and mixed with highly purified water and an organic solvent (3.5 ml of acetonitrile/acetone 30:70 m%) respectively. For the sublimation process, the mixture of DAN and organic solvent is placed in a preheated plate up to 125 degrees to allow the homogenous crystallization of the matrix. At the same time, the ITO slides are placed onto a cooler path of about 10 degrees. Both the mixture and ITO slides are then placed into a high vacuum-sealed chamber with approximately 5.2x 10^{-2} mbar pressure for around 15 to 20 min. Then, the matrix evaporates towards the sample slides and forms a thin and even layer shown in Figure 27: DAN matrix sublimation. The ITO coated slides are weighed before and after the sublimation in order to know how much matrix amounts are placed on the slides. Table 1 and Table 2 show the slide weights before and after sublimation for the Austrian and Syrian donors respectively, before and after caffeine consumption.

Volunteers	ITO slides with Caffeine		ITO slides without Caffeine	
	Slides without Matrix [mg]	Slides with Matrix [mg]	Slides without Matrix [mg]	Slides with Matrix [mg]
Volunteer 1	4917,2	4921,1	4910,0	4914,1
Volunteer 2	4909,9	4913,8	4908,4	4912,2
Volunteer 3	4909,2	4914,3	4882,9	4887,1

Table 1: Austrian volunteers experiment slides weight in the presence and absence of caffeine before and after sublimation.

[3] https://www.nespresso.com/at/en/

Volunteers	ITO slides with Caffeine		ITO slides without Caffeine	
	Slides without Matrix [mg]	Slides with Matrix [mg]	Slides without Matrix [mg]	Slides with Matrix [mg]
Volunteer 1	4922,3	4925,8	4914,9	4918,7
Volunteer 2	4916,3	4920,7	4919,3	4923,3
Volunteer 3	4921,0	4925,1	4925,4	4928,6

Table 2: Syrian volunteers experiment slides weight in the presence and absence of caffeine before and after sublimation.

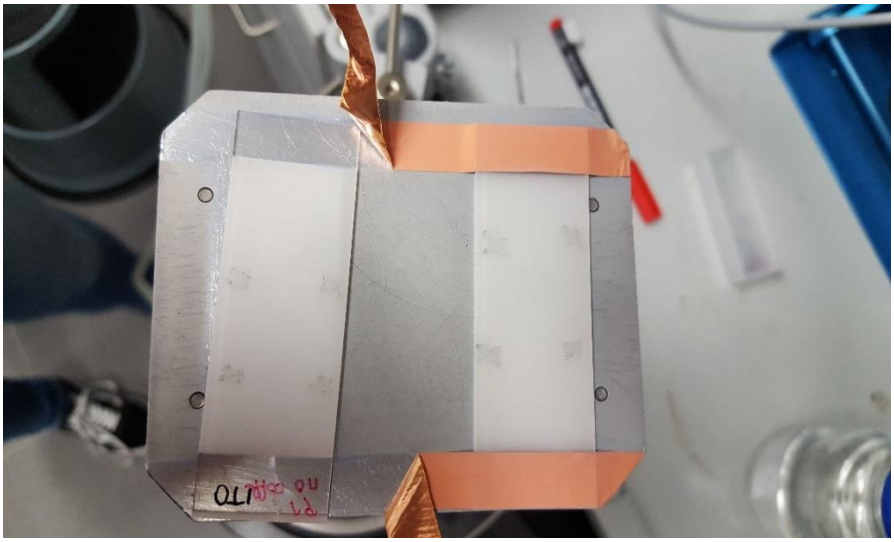


Figure 27: DAN matrix sublimation. The DAN is briefly heated for crystallization and is placed into a vacuum-sealed chamber with the sample slides. This forms a thin and even layer of crystals on the sample surface.

The next step is the optical scanning. The slides are entered into image scanner similar to the normal scanners used to convert the fingerprint slides to a digitalized image as shown in Figure 28.

Figure 28: Scanning of fingerprint slides to obtain digitalized fingerprint images.

5.1.2 The measurement and calibration process in MALDI MS

In order to extract an accurate mass spectrum of our analyte of interest, an internal mass calibration in MALDI is required by using several specific calibrant signals and known peaks. Otherwise there will be a spectrum misalignment and some errors in analytes identification will happen [129]. This step increases the calibration precision standards and reduces the incorrect evaluation of our data. For this purpose, red phosphorus is used since it forms distinct monoisotopic clusters that are nontoxic, stable, amorphous, and suitable for lipid imaging [130]. For the mass calibration, roughly 3 mg of red phosphorus is dissolved in 1 ml of ultra-high-quality water. Then, a few drops of the solution are placed on the slides away from the fingermark area since red phosphorus has known molecular weight (30.974 g/mol) and covers the range up to approximately 3000 m/z in both negative and positive ion modes.

After the calibration process, the samples are ready to be measured. Fingerprint samples imaging and profiling were done using an Ultra-flex MALDI pulsed ultraviolet laser at wavelength 337 nm within a 100µm raster step [131] (raster width is the distance between two shots and it indicates the image spatial resolution) and can produce around 300 laser shots per raster step using a scan range from 1 to 1000 Da. Fingerprint samples imaging and profiling were done using an Ultra-flex MALDI pulsed ultraviolet laser at wavelength 337 nm within a 100µm raster step [131] (raster width is the distance between two shots and it indicates the image spatial resolution) and can produce around 300 laser shots per raster step using a scan range from 1 to 1000 Da. It takes three to five hours to irradiate one fingerprint in one mode. Since we have 24 fingerprints in total, a 5x5 mm square from the original image only is selected to reduce time needed for measurement. The laser irradiates the pre-defined sample area for each single pixel with 2500 positions and the blank square (as a reference to discriminate between matrix and analyte) in positive and negative modes.

5.2 Data acquisition problems and preprocessing procedures

In the theoretical MALDI principle, the sample is exposed to laser shots producing molecular ions hitting the detector during specific times called time resolution of the instrument. The whole spectrum is obtained within a few nanoseconds and contains thousands of m/z values. However, these spectral data contain information about the sample as well as noises and artifacts which are originated from several sources. The ions peaks may also shift, causing some error in the experiments. This produces inaccurate spectral data and incorrect identification of the components in the sample. Thus, we need to apply essential preprocessing algorithms to correct them and obtain accurate statistical analysis of the data. ImageLab software is implemented to import and store these data as a hypercube for all data sets. The axes of the hypercube are defined as following: X (horizontal spatial coordinate), Y (vertical coordinate), L (layers), and T (time slot). Some preprocessing procedures (shown in Figure 29) are performed in order to enhance the hyperspectral images and overcome the low sensitivity and acquisition problems of the instrument. These steps and the reasons behind them are explained in later in this work. The total number of datasets is 22 fingerprints instead of 24 because we had a problem in exporting the rest two files from the instrument PC for unknown reasons.

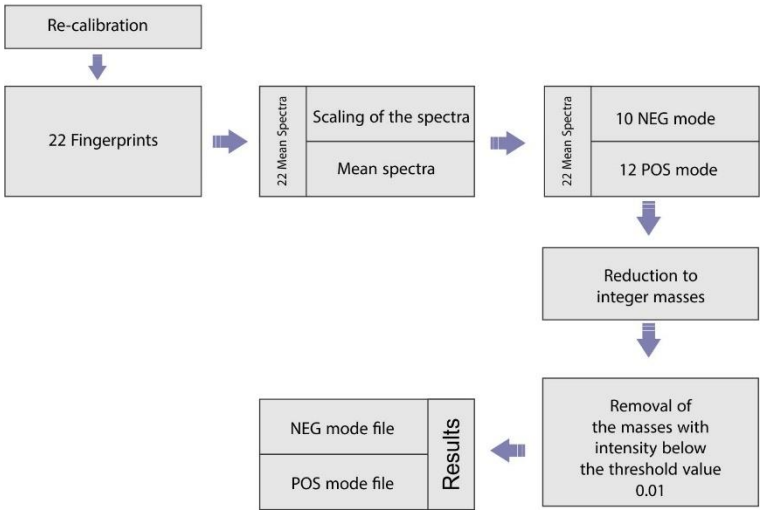


Figure 29: Preprocessing procedures to enhance the data and overcome the acquisition problems.

1- Recalibration of mass spectra

Since the used MALDI instrument doesn't support calibrated data export (calibration here means to match the molecular ions time of flight with their m/z values), we manually apply several calibration points to each image (ImageLab: Tool-> recalibration-> recalibrate spectra) as shown in Figure 30.

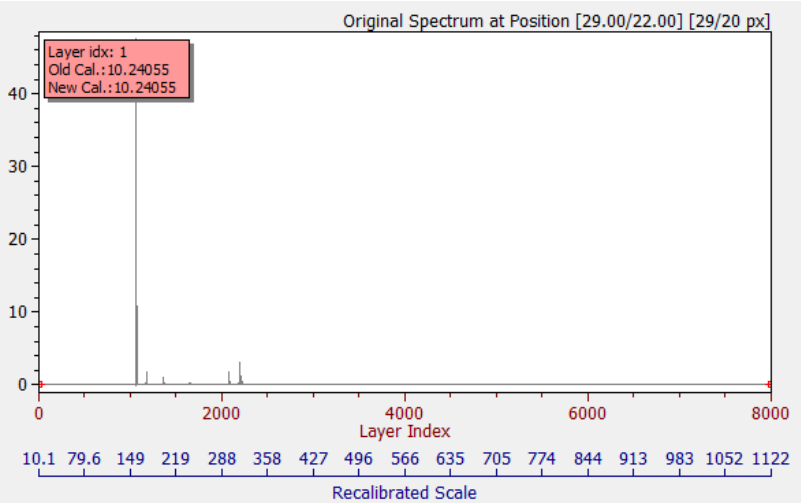


Figure 30: Recalibration of the spectrum

2- Trimming the data matrix

As mentioned previously, around 2500 positions and the blank square in each image for each mode (positive and negative) are irradiated by the UV laser. However, some deformations in the image have emerged during the acquisition process and can cause some misalignments as shown in Figure 31 (left). The reason behind this error is that the laser was not correctly adjusted at the beginning of the square, so the shooting process didn't not start from the very first pixel, causing a deformation of the desired image. Moreover, the image also contains the blank values and zero values (background) which don't hold any information about our sample and take up time and storage space in the data cube. Thus, the excluding procedure is important in this case to get rid of all the non-important

information. This is done by trimming the data matrix as shown in Figure 31 (b) i.e. the minimum and maximum index are selected for each X and Y dimension in each image while L and T stay the same (ImageLab: Tools -> process raw data ->Trim data matrix). After trimming the data matrix, all the data outside the defined range are removed.

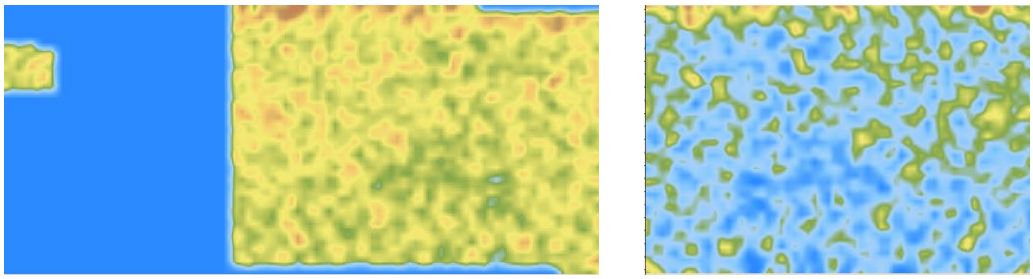


*Figure 31: Trimming the data cube: **Left**: before the trimming, **Right**: after the trimming.*

3- Scaling and averaging the mass spectra

In order classify the "CF/non-CF" and "Austrian/Syrian" groups, a direct comparison between the spectra is not effective due to the ion intensities' variation from spectrum to spectrum within the same sample. Thus, all the ion intensities of the spectra should be scaled to a common scale to minimize these differences which result from several sources including sample preparation and experimental errors (i.e. we can't achieve the optimal fingerprint acquisition due to some uncontrolled errors like unbalanced compounds distribution of the sample due to unbalanced pressure patterns of the donator index finger). One of the simplest methods to facilitate the direct comparison of the samples without losing information at the same time is scaling to the constant sum. This means a constant number of units are applied to each feature of the target. Scaling can be done automatically via ImageLab software by summing up the values of all intensity peaks in each pixel and then calculating the conversion factor k. This factor converts the sum of the whole spectrum to a constant value and then multiplies it by k. Thus, a new scaled spectrum is achieved (ImageLab: Preprocessing -> scaling the data to constant sum).

Because the finger ridges are not clearly visible in our images (due to low spatial resolution acquisition of the slides that resulted from preparation/instrument problems), the average spectrum for each image is calculated to overcome this problem. This can be done by summing up all spectra values for one image, then dividing the result by the number of spectra. Furthermore, the standard deviation for each mean spectrum is calculated to provide some statistical information about the data.

Finally, we collected 22 mean spectra which can be categorized in the following table (there are two missing spectra because of two corrupted files due to an instrument software failure during the exporting phase).

Mass spectrum	CF/non-CF	Country of origin	Mode
Spectrum 1	CF	Austrian	Negative
Spectrum 2	non-CF	Austrian	Negative
Spectrum 3	CF	Austrian	Negative
Spectrum 4	non-CF	Austrian	Negative
Spectrum 5	CF	Austrian	Negative
Spectrum 6	non-CF	Syrian	Negative
Spectrum 7	CF	Syrian	Negative

Spectrum 8	non-CF	Syrian	Negative
Spectrum 9	CF	Syrian	Negative
Spectrum 10	non-CF	Syrian	Negative
Spectrum 11	CF	Austrian	Positive
Spectrum 12	non-CF	Austrian	Positive
Spectrum 13	CF	Austrian	Positive
Spectrum 14	non-CF	Austrian	Positive
Spectrum 15	CF	Austrian	Positive
Spectrum 16	non-CF	Austrian	Positive
Spectrum 17	CF	Syrian	Positive
Spectrum 18	non-CF	Syrian	Positive
Spectrum 19	CF	Syrian	Positive
Spectrum 20	non-CF	Syrian	Positive
Spectrum 21	CF	Syrian	Positive
Spectrum 22	non-CF	Syrian	Positive

Table 3: Table of resulted spectra for each volunteer in positive and negative detection mode for caffeine and non-caffeine consumption.

4- Reduction to the integer masses and removing the masses below specific threshold

Our data set contains thousands of spectra for each fingerprint image, and each spectrum in this image includes several m/z values that corresponds to CF molecules. In theory, those peaks should be located at the exact same position for all spectra because they indicate the same fragment. But due to inaccurate internal mass calibration in MALDI MS (approximately ± 0.1 Da), we have peaks variation of roughly 10 masses, causing all the individual peaks to shift around 10 of the masses. This is problematic because when we sum up all the spectra, those shifted peaks will give us a broader peak and low intensity, and eventually wrong identification of our analyte of interest. To align the peaks and improve the resolution, a digitalization process is used by converting all the peaks to integer masses (i.e. summing up all their intensities and divide them by a range of ± o.5). The final result is integer masses that correspond to our analyte peaks and facilitate the comparison step between “CF/non-CF” and “Austrian/Syrian” groups.

The resulting spectrum contains many peaks. Some of them indicate important information about the sample, but others are noise peaks that don’t hold any information and can negatively affect the analysis precision. Since around 1% of the most intense and important peaks are the target, a simple program was executed using Pascal programming language to discard all the peaks that were below a specific threshold (0.01).

After these preprocessing steps, the resulting data are grouped in two files according to the detection mode (positive and negative) and imported to DataLab software as ASC files. The positive file contains around 25 variables, while the negative file contains around 447 variables. There is a difference between the variable numbers in the positive and negative modes because the lipids and fatty acids are detected and shown clearly in negative mode. Each variable represents an integer mass with its relative intensity value. Several classification models are applied and the results of these models are compared to evaluate the best method for our datasets.

6 Results

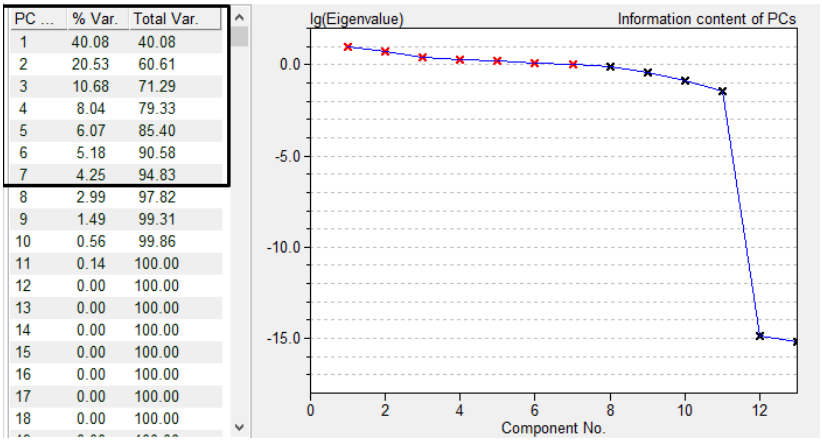
6.1 Caffeine and non-caffeine consumption discrimination

6.1.1 Positive detection mode

To differentiate between CF and non-CF in positive detection mode. The CF class is coded by integer 1 and non-CF class is coded by integer 0. The number of independent variables in positive mode are 12 samples and around 25 dependent variables.

6.1.1.1 Principal component analysis

In order to calculate the principal components of our variables set (25 variables), standardized scaling is carried out since it scales all variables to a unified scale so we can see comparable values. Figure 32 depicts PCA results of the data showing that the first 7 variables are sufficient, since they cover more than 90% (around 94.83%) of the total variance.



PC ...	% Var.	Total Var.
1	40.08	40.08
2	20.53	60.61
3	10.68	71.29
4	8.04	79.33
5	6.07	85.40
6	5.18	90.58
7	4.25	94.83
8	2.99	97.82
9	1.49	99.31
10	0.56	99.86
11	0.14	100.00
12	0.00	100.00
13	0.00	100.00
14	0.00	100.00
15	0.00	100.00
16	0.00	100.00
17	0.00	100.00
18	0.00	100.00

Figure 32: The result of calculating principal component analysis of variables set. This result shows that the first 7 variables are the most important ones which contains the most valuable information.

Then the first components are used to make regression analysis (which means PCR).

```
Standard Dev. of Residuals ....: 0.4556
Quality of Fit ................: 0.7232
F-Statistic ...................: 1.493
Mean of Target Values .........: 0.500000
Std.Dev. of Target Values .....: 0.522233
Mean of Calculated Values .....: 0.500000
Std.Dev. of Calc. Values ......: 0.444127

Regression coefficients for principal components:
      PC          Coefficient       StdDev(coeff)   t-value  alpha
   ------------------------------------------------------------------
    INTERCEPT     5.0000000E-01 +/- 1.3151884E-01      3.802  0.0191
   1 [ 40.08 %]  -5.7320974E-02 +/- 4.3395980E-02     -1.321  0.2570
   2 [ 20.53 %]   8.4404427E-02 +/- 6.0626887E-02      1.392  0.2363
   3 [ 10.68 %]   1.3863299E-01 +/- 8.4086632E-02      1.649  0.1746
   4 [  8.04 %]   6.9951600E-03 +/- 9.6902991E-02      0.072  0.9459
   5 [  6.07 %]   9.1188625E-02 +/- 1.1149688E-01      0.818  0.4594
   6 [  5.18 %]  -2.4383842E-02 +/- 1.2074384E-01     -0.202  0.8498
   7 [  4.25 %]  -2.4339135E-01 +/- 1.3323380E-01     -1.827  0.1418
```

Figure 33: Principal compoenents regression coefficients.

6.1.1.2 *Variable selection and multiple linear regression*

The idea behind MLR model is to predict the relationship between dependent variables and independent variables. For this purpose, the selection of variables is essential for testing the descriptors and increase the adequacy of the analyzing process. DataLab includes variable selection tool. The working principle is described in Chapter 3.

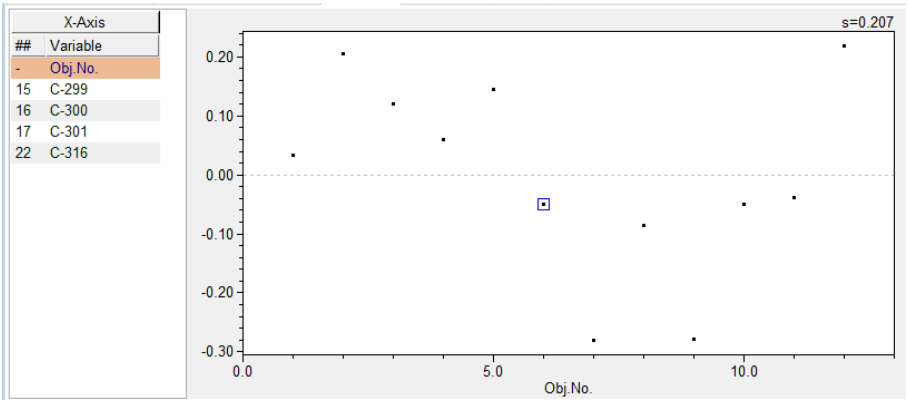
Target Variable: CF/non-CF

Selected Models

Model Variables	RMS	min|t|	AIC	BIC	F	r2
18	0.33688	3.929	-25.16	-24.67	14.03	0.5839
18, 16	0.33027	1.202	-24.78	-23.81	7.877	0.6364
18, 16, 17	0.31242	1.475	-25.37	-23.92	6.440	0.7072
18, 16, 17, 15	0.27734	1.469	-27.65	-25.71	6.782	0.7949
18, 16, 17, 15, 22	0.20698	0.01961	-34.27	-31.85	10.80	0.9000
16, 17, 15, 22	0.19362	3.584	-36.27	-34.33	15.76	0.9000
16, 17, 15, 22, 13	0.13117	3.229	-45.22	-42.79	28.69	0.9599
16, 17, 15, 22, 13, 21	0.072342	4.125	-59.35	-56.44	78.78	0.9895
16, 17, 15, 22, 13, 21, 20	0.030900	5.281	-79.95	-76.56	358.5	0.9984
16, 17, 15, 22, 13, 21, 20, 5	0.025390	1.845	-85.34	-81.47	435.9	0.9991
16, 17, 15, 22, 13, 21, 20, 5, 25	0.023115	1.351	-89.05	-84.69	415.7	0.9995

Figure 34: Variable selection for the CF/non-CF target variable. The first four variables are selected to be introduced to MLR model.

Stepwise regression is selected and all variables (25 [m/z] values) are fed to the selection tool. In fact, due to the low number of samples and to avoid getting a random model, the best four variables that have higher F and R^2 values were chosen, according to rule of thumb in MLR model which states that the number of selected variables should be roughly the third of number of objects (we have 12 negative mode objects which leads to 4 variables). The fourth selected variables (16, 17, 15, 22) have mass values (300, 301, 299, 316) respectively. Finally, the MLR model is calculated based on the selected descriptors variables.



Regression coefficients:

Col-#	Var-Name	Coefficient	Std.Err.(coeff)	t-Test	alpha
-	INTERCEPT	-7.3962796E-01	+/- 2.6450331E-01	-2.796	0.0267
15	C-299	8.0452969E-01	+/- 1.3067369E-01	6.157	0.0005
16	C-300	-9.7081652E-01	+/- 2.2753153E-01	-4.267	0.0037
17	C-301	2.2765129E+00	+/- 3.4739664E-01	6.553	0.0003
22	C-316	-2.3718322E-01	+/- 7.0754404E-02	-3.352	0.0122

*Figure 35: Shows multivaraiate regression of positive detection mode dataset: **top**) Residual plot and **bottom**) Regression coefficients.*

At first glance, the resulting MLR model shows a good prediction (high R^2 and F values), which means the linear regression hyper plan can adapt adequately through the selected points. This can be shown obviously in residual plot which illustrate very good MLR model performance for CF and non-CF classes.

These good results can be partially justified by overfitting, since MLR model overestimates the precision of the model in case of a low number of observations where erroneous data could be fitted by the model, then the acquired model prediction meaningless.

We can also notice here that (300, 301, 299, 316 [m/z]) values don't correspond to caffeine peaks or either of its metabolites. Therefore, this model is not beneficial for discrimination based on caffeine consumption, but instead is high likely that it formed based on unrelated arbitrary data.

6.1.1.3 Partial least squares-based discriminant analysis (PLS/DA)

PLS/DA tool in DataLab is used. It creates a model to discriminate between two groups using all variables to build a classifier.

The number of factors can correspond to the number of objects and standardization scaling mode based on the correlation matrix is applied on the variables. Figure 36 shows the separation between two classes: red and blue which indicate CF and non-CF respectively. The classification results are presented as a confusion matrix where the green and gray colors indicate the true positive (TP) and true negative (TN) respectively. The confusion matrix describes the classification between two classes, therefore, a significant prediction of the spectral data is achieved. This can be seen as a result of low number of samples which means fitting the erroneous data.

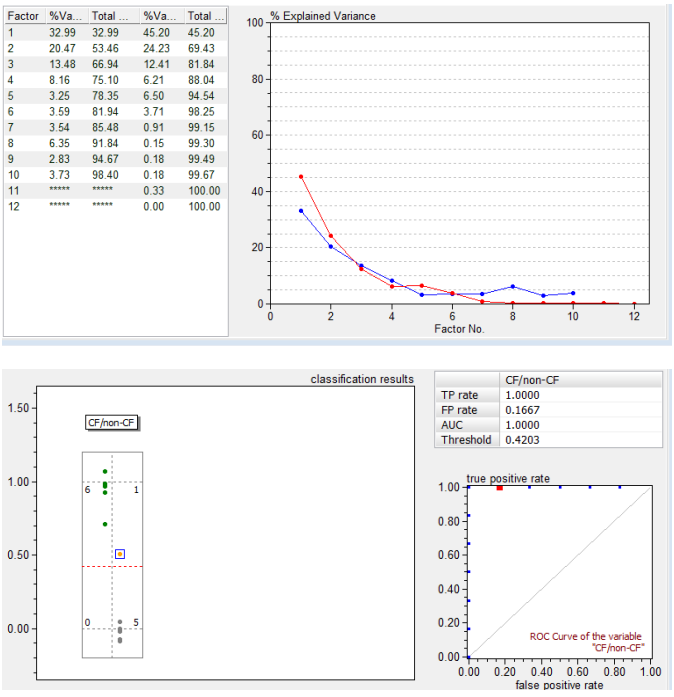


*Figure 36: **Top**) results of PLSDA classifier of CF and non-CF classes and list of variables. **Bottom**) shows the model performance based on the confusion matrix showing how many objects are correctly classified. The objects are shown as green points in the true positive column for CF class and gray points in true negative column for non-CF class.*

To estimate the best number of factors for PLSDA model, cross validation is performed to analyze the training set and then validate this analysis on the test set by choosing the appropriate test set size and the number of repetitions. We can see that to achieve lowest RMS, the best factor number is 4.

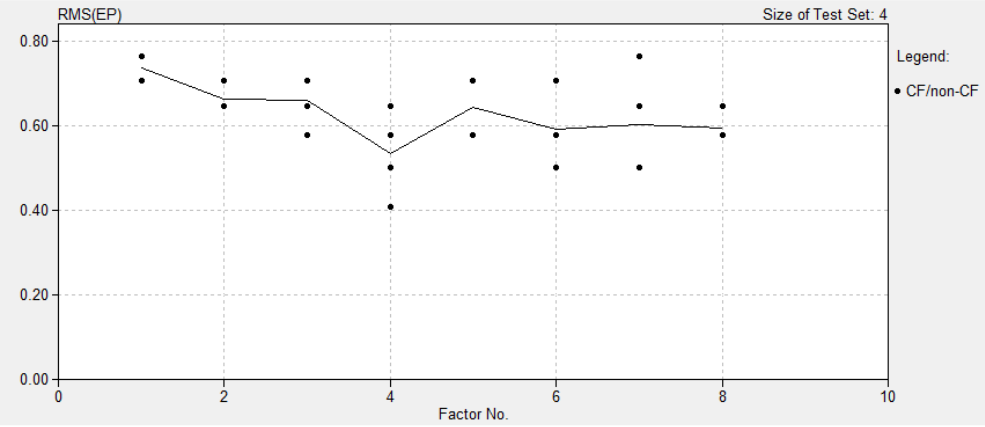


Figure 37: Cross validation for the PLSDA model to evaluate the performance shows the size of the test set (1/4) and number of repetition (3). Least error model is the one with 4 factors.

6.1.1.4 Compare PCR, MLR and PLS models and discussion

Assuming Bessel correction, we can assume that standard deviation of the residuals can correspond to standard deviation of the error in previous models. PLS has an error standard deviation about 0.0717, while MLR and PCR have about 0.2070 and 0.4885 respectively of residuals standard deviation. We can assume the superiority of PLS on other models. That comes from the fact that PCA usually searches for maximum variance and MLR for best correlation, but PLS for both by seeking maximum covariance between target and descriptors.

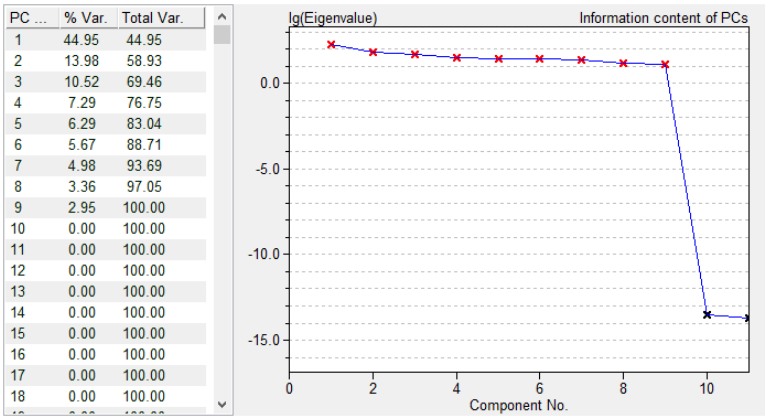
6.1.2 Negative detection mode

The same steps that are described in (7.1) will apply to differentiate between CF and non-CF in negative detection mode. The CF class is coded by integer 1 and non-CF class is coded by integer 0. The number of independent variables in negative mode 12 samples and around 447 dependent variables.

6.1.2.1 Principle component analysis

```
Regression coefficients for principal components:
       PC          Coefficient      StdDev(coeff)    t-value  alpha
   ----------------------------------------------------------------
    INTERCEPT     5.0000000E-01 +/- 1.6007959E-02     31.234  0.0010
   1 [ 46.39 %]   7.4593317E-02 +/- 3.8692835E-03     19.278  0.0027
   2 [ 15.03 %]  -1.9894389E-03 +/- 6.7972289E-03     -0.293  0.7973
   3 [ 10.62 %]   1.4256610E-01 +/- 8.0872322E-03     17.629  0.0032
   4 [  7.91 %]  -7.8725113E-03 +/- 9.3678819E-03     -0.840  0.4892
   5 [  5.28 %]   2.0337355E-02 +/- 1.1472104E-02      1.773  0.2183
   6 [  4.77 %]   7.1841147E-03 +/- 1.2060179E-02      0.596  0.6118
   7 [  4.03 %]  -2.2229453E-01 +/- 1.3124311E-02    -16.938  0.0035
```

Figure 38 shows that the first eight components cover more than 90% (around 93.69%) of the information. We can use those component and apply regression analysis.



PC ...	% Var.	Total Var.
1	44.95	44.95
2	13.98	58.93
3	10.52	69.46
4	7.29	76.75
5	6.29	83.04
6	5.67	88.71
7	4.98	93.69
8	3.36	97.05
9	2.95	100.00
10	0.00	100.00
11	0.00	100.00
12	0.00	100.00
13	0.00	100.00
14	0.00	100.00
15	0.00	100.00
16	0.00	100.00
17	0.00	100.00
18	0.00	100.00

Regression coefficients for principal components:

PC	Coefficient	StdDev(coeff)	t-value	alpha
INTERCEPT	5.0000000E-01	+/- 1.6007959E-02	31.234	0.0010
1 [46.39 %]	7.4593317E-02	+/- 3.8692835E-03	19.278	0.0027
2 [15.03 %]	-1.9894389E-03	+/- 6.7972289E-03	-0.293	0.7973
3 [10.62 %]	1.4256610E-01	+/- 8.0872322E-03	17.629	0.0032
4 [7.91 %]	-7.8725113E-03	+/- 9.3678819E-03	-0.840	0.4892
5 [5.28 %]	2.0337355E-02	+/- 1.1472104E-02	1.773	0.2183
6 [4.77 %]	7.1841147E-03	+/- 1.2060179E-02	0.596	0.6118
7 [4.03 %]	-2.2229453E-01	+/- 1.3124311E-02	-16.938	0.0035

*Figure 38: **Top**) PCA results for the fingerprint data in negative detection mode. It shows that the first 7 components are the most valuable ones which holds the most information. **Bottom**) The coefficient of PCR.*

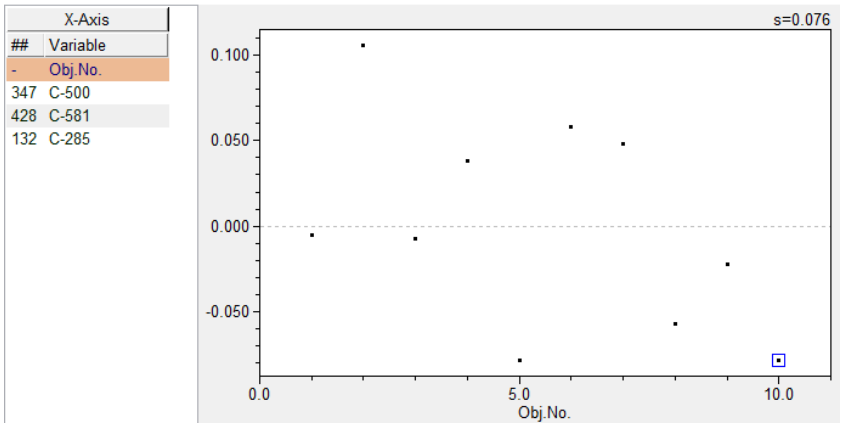
6.1.2.2 Variable selection and multiple linear regression

The number of the objects in the negative detection mode 10 samples. According to the rule that is described in (7.1), the model with three variables is chosen to be introduced to calculate MLR classifier. The three selected variables (347, 428, 132) have the m/z values (500, 581, 285) respectively as shown in next figure.

Target Variable: CF/non-CF

Selected Models

Model Variables	RMS	min|t|	AIC	BIC	F	r2
347	0.18973	7.775	-32.30	-31.99	53.73	0.8704
347, 428	0.10252	4.777	-43.79	-43.18	100.6	0.9664
347, 428, 132	0.070381	3.158	-50.64	-49.74	142.2	0.9861
347, 428, 132, 328	0.032859	5.110	-65.42	-64.21	481.1	0.9974
347, 428, 132, 328, 180	0.011159	6.857	-86.84	-85.33	3211	0.9998
347, 428, 132, 328, 180, 317	0.0026167	9.324	-116.1	-114.3	45640	1.0000
347, 428, 132, 328, 180, 317, 447	1.3503E-04	38.72	-176.2	-174.1	13058045	1.0000
347, 428, 132, 328, 180, 317, 447, 131	5.4166E-07	431.8	-288.7	-286.2	99999999	1.0000
347, 428, 132, 328, 180, 317, 447, 131, 211	1.2059E-15	635221684	-692.1	-689.3	99999999	1.000



Regression coefficients:

Col-#	Var-Name	Coefficient	Std.Err.(coeff)	t-Test	alpha
-	INTERCEPT	-1.0342954E+00 +/-	1.2695464E-01	-8.147	0.0002
347	C-500	1.5959258E+01 +/-	8.8625795E-01	18.007	0.0000
428	C-581	-3.4296813E+00 +/-	5.1300614E-01	-6.685	0.0005
132	C-285	-7.3310649E-02 +/-	2.5071720E-02	-2.924	0.0265

*Figure 39: Shows multivaraiate regression of negative detection mode dataset. **Top**) Variable selection using stepwise regression method. The best model is selected which consists of three variables (347, 428, 132). **Middle**) Residuals plot. **Bottom**) Regression coefficients.*

After that, MLR model is calculated. The result shows good separation between CF and non-CF groups as shown in residuals plot.

6.1.2.3 *Partial least squares-based discriminant analysis (PLS/DA)*

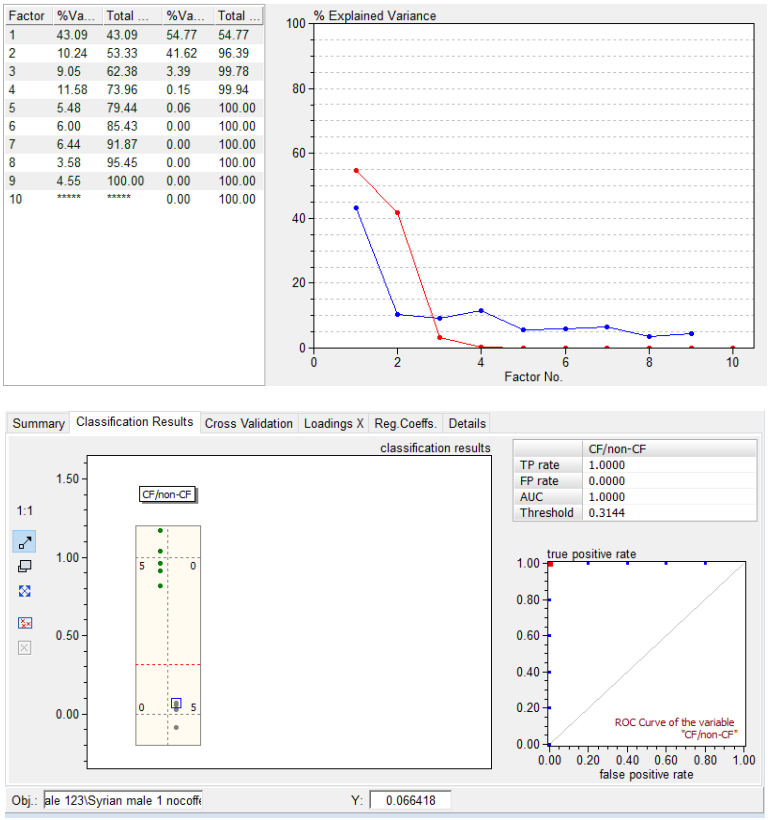


*Figure 40: **Top**) PLS-DA classification results which shows good classification between the CF and non-CF classes in negative detection mode**. Bottom**) The confusion matrix illustrates the validation of our model and he ROC curve in the lower left of the figure shows that the classifier is perfect which is represented in the red point.*

We can see also good model results with 3 factors.

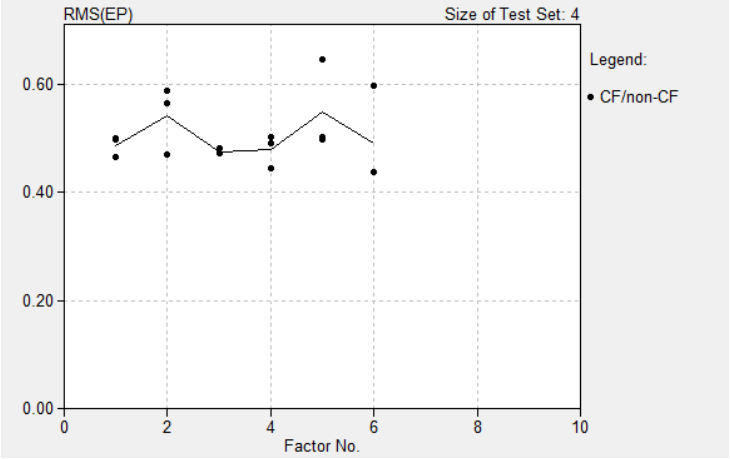


Figure 41: Cross-validation results of the applied model. The size of the test set is ¼ of the whole dataset, the repetition number is 2, and we can see best model contains 3 factors.

6.1.2.4 *Compare PCR, MLR and PLS models and discussion*

Assuming Bessel correction, PLS has an error standard deviation about 0.0054, while MLR and PCR have about 0.2070 and 0.0760 respectively of residuals standard deviation. We can assume the superiority of PLS.

6.1.3 Results discussion

All applied classification models on the fingerprint datasets showed good separation between the two classes (CF/non-CF) in both detection modes (negative/positive), However, PLS model was the best.

Due to the low number of observations, classification algorithms suffered from overfitting problem and we could assume that this model generated randomly by fitting noisy data. This assumption can be supported by looking at variables selection results and t-test results.

Variable selection results showed variables different than the known caffeine and/or metabolites mass to charge ratio values. T-test of caffeine/non caffeine subset also showed no significant difference.

Positive mode:

$$|\mathrm{t}_{\mathrm{statistic}}| = 0.0229 < t_{\alpha=0.05/2} = 1.9680$$

Negative mode:

$$|\mathrm{t}_{\mathrm{statistic}}| = 0.0489 < t_{\alpha=0.05/2} = 1.9605$$

We could also argue that peaks corresponding to caffeine didn't appear clearly in the primary spectra, the thing that prevented us from reaching clear results. The reason behind this could be the volatile nature of caffeine which reduced the chance to obtain good samples, especially if we take into account the problems we faced during experiment preparation and instrument calibration. It could also be the case that the caffeine metabolism might need longer than 3.5 hours to be traceable on the fingerprint in different human bodies.

6.2 Individuals' country of origin discrimination based on lipids variation on fingerprints.

The second task in this work is to classify the fingerprint data sets into Austrian and Syrian groups for positive and negative modes by calculating the described methods in (7.1 and 7.2) sections. The Austrian class is coded by integer 1 and the Syrian is coded by integer 0.

6.2.1 Positive detection mode

6.2.1.1 Principal component analysis

Since the data is the same as presented in 6.1.1.1 , the same steps (Figure 32) are followed, resulting in the following regression coefficients.

```
Standard Dev. of Residuals ....: 0.3270
Quality of Fit ................: 0.8574
F-Statistic ...................: 3.437
Mean of Target Values .........: 0.500000
Std.Dev. of Target Values .....: 0.522233
Mean of Calculated Values .....: 0.500000
Std.Dev. of Calc. Values ......: 0.483580

Regression coefficients for principal components:
      PC          Coefficient       StdDev(coeff)    t-value  alpha
   ------------------------------------------------------------------
    INTERCEPT     5.0000000E-01 +/- 9.4389751E-02      5.297  0.0061
   1 [ 40.08 %]   1.4988414E-02 +/- 3.1144859E-02      0.481  0.6555
   2 [ 20.53 %]   1.7031659E-01 +/- 4.3511308E-02      3.914  0.0173
   3 [ 10.68 %]  -2.9849701E-02 +/- 6.0348131E-02     -0.495  0.6468
   4 [  8.04 %]  -7.6780846E-02 +/- 6.9546303E-02     -1.104  0.3315
   5 [  6.07 %]  -1.4538005E-01 +/- 8.0020189E-02     -1.817  0.1434
   6 [  5.18 %]  -2.3648288E-02 +/- 8.6656638E-02     -0.273  0.7984
   7 [  4.25 %]   1.8313386E-01 +/- 9.5620558E-02      1.915  0.1280
```

Figure 42: Principal compoenents regression model coefficients for discrimination between Austrian and Syrian in positive detection mode.

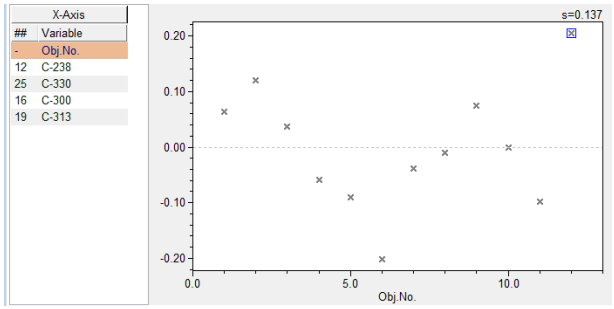
6.2.1.2 Variable selection and multiple linear regression

The four selected variables are shown in the next figure.

The four selected variables (12, 25, 16 and 19) have mass to charge ratio values (238, 330, 300, 313 [m/z]) respectively. Finally, the MLR model is calculated based on the selected descriptors variables.

Target Variable: SYR/AUS

Selected Models						
Model Variables	RMS	min|t|	AIC	BIC	F	r2
12	0.2776	5.284	-29.80	-29.32	25.39	0.7174
12, 25	0.2372	2.249	-32.71	-31.74	19.48	0.8124
12, 25, 16	0.1859	2.698	-37.83	-36.37	23.04	0.8963
12, 25, 16, 19	0.1280	3.312	-46.19	-44.25	38.25	0.9563
12, 25, 16, 19, 10	0.1035	2.289	-50.89	-48.47	46.76	0.9750
12, 25, 16, 19, 10, 8	0.0...	3.249	-61.07	-58.16	91.09	0.9909
12, 25, 16, 19, 10, 8, 17	0.0...	4.697	-79.34	-75.94	340.6	0.9983
12, 25, 16, 19, 10, 8, 17, 2	0.0...	4.340	-98.24	-94.36	1278	0.9997
12, 25, 16, 19, 10, 8, 17, 2, 18	0.0...	4.670	-121.6	-117.2	6262	1.0000
12, 25, 16, 19, 10, 8, 17, 2,...	0.0...	7.443	-159.9	-155.0	80877	1.0000
12, 25, 16, 19, 10, 8, 17, 2,...	3.1...	6....	-754.2	-748.9	99...	1.000




```
Regression coefficients:
 Col-#     Var-Name       Coefficient       Std.Err.(coeff)    t-Test   alpha
 ------------------------------------------------------------------------------
     -   INTERCEPT      1.8669693E-01 +/- 1.4792750E-01      1.262   0.2473
    12   C-238          2.8144360E+00 +/- 2.4580731E-01     11.450   0.0000
    25   C-330         -6.0749554E+00 +/- 1.1886136E+00     -5.111   0.0014
    16   C-300          5.1031793E-01 +/- 1.1326485E-01      4.506   0.0028
    19   C-313         -1.5932754E+00 +/- 5.1430992E-01     -3.098   0.0174
```

*Figure 43: Shows multivaraiate regression of positive detection mode dataset. **Top**) Variable selection for the Syrian/Austria target variable. The first four variables are selected to be introduced to MLR model. **Middle**) Residual plot. **Bottom**) Regression coefficients.*

6.2.1.3 Partial least squares-based discriminant analysis (PLS/DA)

Figure 36 shows the separation between two classes, red and blue, which indicate Syrian and Austrian respectively. The classification results are presented as a confusion matrix where the green and gray colors indicate the true positive (TP) and true negative (TN) respectively. The confusion matrix describes the clear classification between two classes. Therefore, a significant prediction of the spectral data is achieved. This can be seen as a result of low number of samples which means fitting the erroneous data.

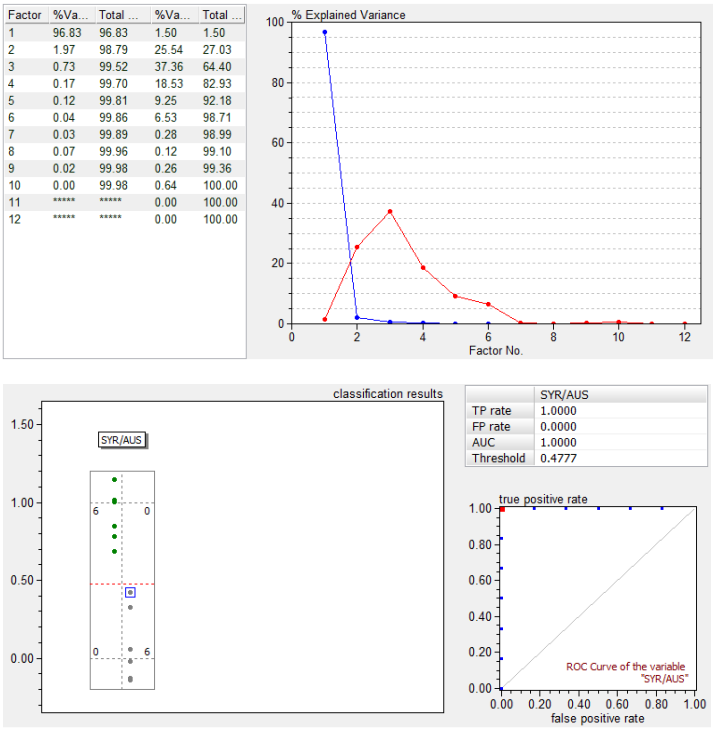


*Figure 44: **Top**) Results of PLSDA classifier of CF and non-CF classes and list of variables. **Bottom**) shows the model performance based on the confusion matrix shows how many objects are correctly classified. The objects are shown as green points in the true positive column for CF class and gray points in true negative column for non-CF class.*

To estimate the best number of factors for PLSDA model, cross validation is performed.

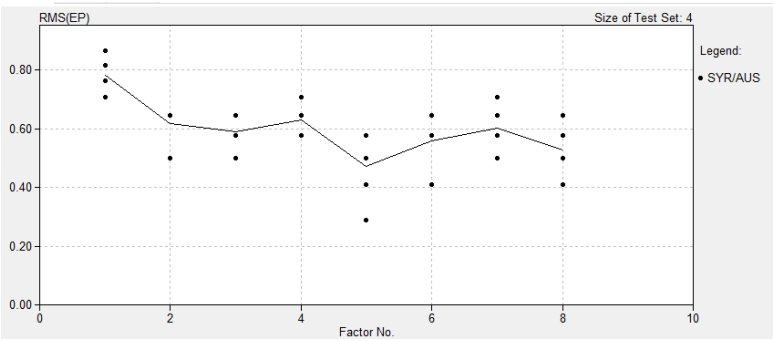


Figure 45: Cross validation for the PLSDA model to evaluate the performance shows the size of the test set (1/4) and number of repetition (5). Least error resulted from a model with 5 factors.

6.2.1.4 Compare PCR, MLR and PLS models

PLS has an error standard deviation about 0.1231, while MLR and PCR have about 0.1369 and 0.3270 respectively of residuals standard deviation. Consequently, PLS is the least error model.

6.2.2 Negative detection mode

6.2.2.1 Principal component analysis

Following same steps as 6.1.2.1, we end up with the following coefficients.

Regression coefficients for principal components:

PC	Coefficient	StdDev(coeff)	t-value	alpha
INTERCEPT	5.0000000E-01	+/- 4.4100981E-02	11.338	0.0077
1 [46.39 %]	1.3327748E-02	+/- 1.0659647E-02	1.250	0.3376
2 [15.03 %]	-5.0531218E-02	+/- 1.8725964E-02	-2.698	0.1143
3 [10.62 %]	-2.1380159E-01	+/- 2.2279847E-02	-9.596	0.0107
4 [7.91 %]	5.5791983E-02	+/- 2.5807961E-02	2.162	0.1632
5 [5.28 %]	5.4901170E-02	+/- 3.1604968E-02	1.737	0.2245
6 [4.77 %]	-2.0544335E-02	+/- 3.3225080E-02	-0.618	0.5994
7 [4.03 %]	-1.5141194E-01	+/- 3.6156702E-02	-4.188	0.0526

Figure 46: Principal components regression model coefficients of Austrian/Syrian based on negative detection mode data.

6.2.2.2 Variable selection and multiple linear regression

For the negative mode file that contains 10 samples, we will select a model that includes 3 variables (15, 60, and 140) with their relative mass to charge ratio values (164, 210, and 293 m/z) respectively as shown in the next figure.

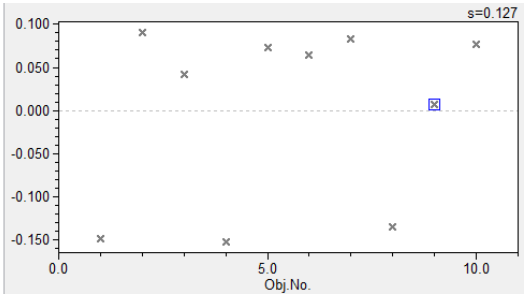
Target Variable: SYR/AUS

Selected Models

Model Variables	RMS	min|t|	AIC	BIC	F	r2
15	0.32766	3.780	-21.37	-21.07	12.70	0.6135
15, 60	0.16669	5.174	-34.06	-33.46	35.86	0.9111
15, 60, 140	0.11730	3.026	-40.43	-39.52	49.91	0.9615
15, 60, 140, 223	0.033673	8.885	-64.93	-63.72	458.1	0.9973
15, 60, 140, 223, 205	0.010157	7.807	-88.72	-87.21	3877	0.9998
15, 60, 140, 223, 205, 175	9.0999E-04	24.88	-137.2	-135.4	377382	1.0000
15, 60, 140, 223, 205, 175, 331	2.4802E-05	73.36	-210.1	-208.0	99999999	1.0000
15, 60, 140, 223, 205, 175, 331, 243	1.1760E-07	365.3	-319.2	-316.8	99999999	1.0000
15, 60, 140, 223, 205, 175, 331, 243, 76	2.1475E-15	77447477	-680.5	-677.8	99999999	1.000

X-Axis

##	Variable
-	Obj.No.
15	C-164
60	C-210
140	C-293



Regression coefficients:

Col-#	Var-Name	Coefficient	Std.Err.(coeff)	t-Test	alpha
-	INTERCEPT	-8.5821349E-01 +/-	2.1824263E-01	-3.932	0.0077
15	C-164	1.7789844E+01 +/-	2.4009070E+00	7.410	0.0003
60	C-210	1.5523146E+01 +/-	2.1938514E+00	7.076	0.0004
140	C-293	-4.6637864E+00 +/-	1.6648523E+00	-2.801	0.0311

*Figure 47: Shows multivaraiate regression of negative detection mode dataset. **Top**) Variable selection for the Syrian/Austria target variable in negative modes. The first three variables are selected to be introduced to MLR model. **Middle**) Residual plot. **Bottom**) Regression coefficients.*

6.2.2.3 Partial least squares-based discriminant analysis (PLS/DA)

As discussed in 6.2.1.3 , the result of PLS DA applied to negative mode dataset are shown below.

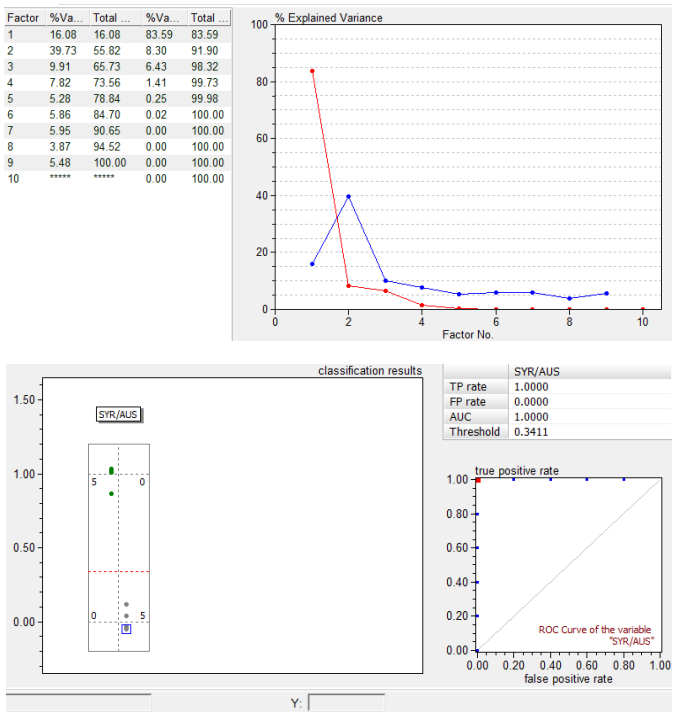


*Figure 48: **Top**) Results of PLSDA classifier of CF and non-CF classes and list of variables**. Bottom**) shows the model performance based on the confusion matrix shows how many objects are correctly classified. The objects are shown as green points in the true positive column for CF class and gray points in true negative column for non-CF class.*

Cross validation showed that 5 factors produce the least error model.

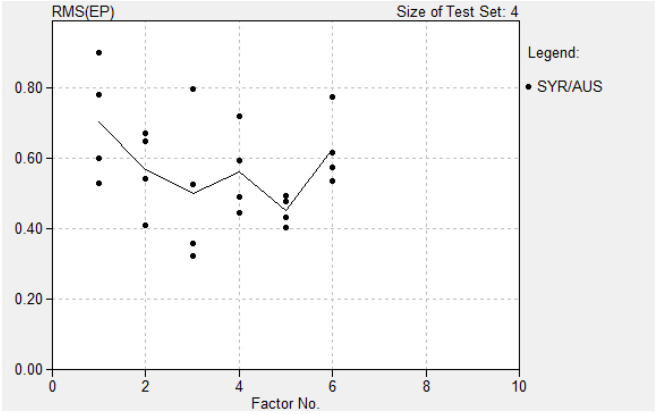


Figure 49: The cross validation results of the applied model, selecting the size of the test set (4) and number of repletion (4) the suitable number of factors is 5 in this case.

6.2.2.4 Compare PCR, MLR and PLS models

PLS has an error standard deviation about 0.0412, while MLR and PCR have about 0.1267 and 0.1797 respectively of residuals standard deviation, so PLS is the least error model again.

6.2.3 Results discussion

Both models in negative and positive modes have shown good separation between Austrian and Syrian, but due to low number of participants, we couldn't assume the correctness of the results. Most likely, these models resulted randomly biased to the data used. This is supported by the results of T-tests. Applying a two-sample two-tailed t-test in positive mode, we can see results where the absolute $|t_{statistics}|$ = (0.0150) is smaller than $t\alpha_{(0.05)/2}$ (1.9680) which means that the null hypothesis can't be rejected. Therefore, no significant difference can be noticed. For negative mode, ($|t_{statistics}|$= 0.365 $<t_{\alpha(0.05)/2}$ = 1.9605) which means that the null hypothesis can't be rejected either.

This also can be seen as a result of not knowing exact features (m/z values) which could make a difference between individuals based on origin. Applying tests on the whole pre-processed spectrum has a high risk of including a huge amount of noise that reduces the effectiveness of the test. However, variable selection results (238, 330, 300, 313) [m/z] in positive mode and (164, 210, and 293) [m/z] in negative mode has potential to be among the values that can help differentiate between individuals based on geographical location. This has a check from the biological perspective to check whether they originated from biological difference (i.e. metabolism, secretion etc.) or merely by chance.

6.3 Other aspects

This experiment has been conducted based on the assumption that all participants are in good health, have no gastrointestinal diseases, and have normal dietary preferences. Any change in the metabolism of the individual can affect the metabolite secretion as well as fingerprint components and intensities. The difference in the intensities between the members of one group (CF/non-CF) or (Austrian/Syrian) can also play a role in reducing the strength of the classifiers.

It is noteworthy to mention that this study can be improved by including a larger number of measurements and participants. Because so much time and effort is involved in mass spectrometry imaging, we tried a kind of tradeoff between the experiment preparation cost and number of measurements. This resulted in the current number of observations.

6.4 Improvement and future work

To check the correctness of the models provided in this thesis, it is recommended to perform the experiment with a larger number of participants (the minimum number of participants depends on the number of features used and classifier model type).

During measurement 5x5 mm squared area is selected to be irradiated and analyzed, this has done to reduce measurements time. Then we could expect more accurate information if we include the whole area of finger print.

It is also recommended to conduct the fingerprint acquisitions at different waiting times to provide a better understanding of the time needed for caffeine to be metabolized and override the possible variation of metabolism time.

Another good suggestion is to extend the participants' group to include both genders. It is tempting in this case to observe whether there is a difference based on gender.

This study used certain classification models (PCR, MLR and PLS DA). It can be a good choice to use additional or different classifiers which may have better performance with hyperspectral datasets.

7 Conclusion

The main aim of this study was to use hyperspectral imaging techniques and MALDI mass spectrometry to differentiate between individuals in cases of caffeine/no caffeine consumption and in cases of different regions of origin from their fingerprints. To achieve this, volunteers from two different countries participated and had their fingerprints collected before and after drinking coffee.

This thesis demonstrated the steps needed to conduct this experiment, showing how the collected fingerprints are converted into different sets of data (negative and positive detection mode) and how data are combined with digital images to form a hypercube of data. Some problems have been raised during experiment preparation and data processing, including: low number of samples, laser adjustment problems in MALDI, the mass calibration applied wrongly for some spectra resulting in peaks shift, an obtaining low resolution images of the samples, hindering the complete usage of the hypercube. To overcome these obstacles, some procedures were applied including: trimming the data hypercube to neglect noise and avoid any misalignment, scaling and averaging the spectra to get rid of low resolution image and dealing solely with spectra to convert information to integer masses, override peaks shift, and apply a threshold in each spectrum to reduce the noise. As a result, two sets of data contain intensities of the average thresholded spectra for each participant were concluded.

To classify each dataset, three classification methods were applied: Principal Component Regression, Multivariate Regression, and Partial Least Squares Discriminant Analysis. All of these methods provided good separation between groups, but PLS was the model which introduced the least error.

Although these classifications were good, due to the issues that have emerged during the experiment (especially low sample numbers), it cannot be proven that these models are correctly classifying the groups. However, it is highly likely that this occurred randomly because of overfitting and noisy, biased data points. This conclusion can be supported by variable selection results which showed different variables/spectral peaks from actual knowledge, and by statistical tests results. A few recommendations were discussed to improve work flow and generate more stable real descriptive models.

References

[1] L. Gu *et al.*, “Fingerprint, Forensic Evidence of,” in *Encyclopedia of Biometrics*, Boston, MA: Springer US, 2009, pp. 528–535.

[2] W. Song *et al.*, “Detection of protein deposition within latent fingerprints by surface-enhanced Raman spectroscopy imaging,” *Nanoscale*, vol. 4, no. 7, p. 2333, Apr. 2012.

[3] M. Asahina, A. Poudel, and S. Hirano, “Sweating on the palm and sole: physiological and clinical relevance,” *Clin. Auton. Res.*, vol. 25, no. 3, pp. 153–159, Jun. 2015.

[4] R. Wadhwa, M. Kaur, and K. V. P. Singh, “Age and Gender Determination from Finger Prints using RVA and dct Coefficients,” 2013.

[5] Z. Zhou and R. N. Zare, “Personal Information from Latent Fingerprints Using Desorption Electrospray Ionization Mass Spectrometry and Machine Learning,” *Anal. Chem.*, vol. 89, no. 2, pp. 1369–1372, Jan. 2017.

[6] M. J. Bailey *et al.*, “Rapid detection of cocaine, benzoylecgonine and methylecgonine in fingerprints using surface mass spectrometry.,” *Analyst*, vol. 140, no. 18, pp. 6254–9, Sep. 2015.

[7] A. H. Chau, J. T. Motz, J. A. Gardecki, S. Waxman, B. E. Bouma, and G. J. Tearney, “Fingerprint and high-wavenumber Raman spectroscopy in a human-swine coronary xenograft in vivo.,” *J. Biomed. Opt.*, vol. 13, no. 4, p. 040501, 2008.

[8] K. C. O’Neill and Y. J. Lee, “Effect of Aging and Surface Interactions on the Diffusion of Endogenous Compounds in Latent Fingerprints Studied by Mass Spectrometry Imaging,” *J. Forensic Sci.*, vol. 63, no. 3, pp. 708–713, May 2018.

[9] A. Van Dam, F. T. Van Beek, M. C. G. Aalders, T. G. Van Leeuwen, and S. A. G. Lambrechts, “CHAPTER 11 COMPARISON OF DIFFERENT TECHNIQUES THAT ACQUIRE DONOR PROFILING INFORMATION FROM FINGERMARKS-A REVIEW.”

[10] Dr. Michael Gozin, “MALDI TOF Imaging of Latent Fingerprints a Novel Bio-signature Tool,” *By Mr. Bogdan Belgorodsky, Dr. Ludmila Fadeev, Dr. Michael Gozin Sch. Chem. Tel Aviv Univ. Tel Aviv 69978,* p. 10, 2010.

[11] I. R. M. Ramos, A. Malkin, and F. M. Lyng, “Current Advances in the Application of Raman Spectroscopy for Molecular Diagnosis of Cervical Cancer,” *Biomed Res. Int.*, vol. 2015, pp. 1–9, 2015.

[12] A. K. Jain, S. Prabhakar, and S. Pankanti, “On the similarity of identical twin fingerprints,” *Pattern Recognit.*, vol. 35, no. 11, pp. 2653–2663, Nov. 2002.

[13] “Can a person fail at biometric fingerprint test if he had injury on thumb ? - Quora.” [Online]. Available: https://www.quora.com/Can-a-person-fail-at-biometric-fingerprint-test-if-he-had-injury-on-thumb. [Accessed: 17-Oct-2018].

[14] R. Ang, X. Yi, and J. Coumbaros, “NOVEL POWDER METHODS FOR THE VISUALIZATION OF LATENT FINGERPRINTS : THE CASE FOR TURMERIC AND OTHER SPICES,” 2018.

[15] J. Eric H. Holder, Laurie O. Robinson, and John H. Laub, *The fingerprint : sourcebook.* U.s. Department of Justice, 2014.

[16] R. Ramotowski, *Lee and Gaensslen’s advances in fingerprint technology, Third Edition - CRC*

Press. CRC Press, 2012.

[17] D. Bovell, "The human eccrine sweat gland: Structure, function and disorders," *J. Local Glob. Heal. Sci.*, vol. 2015, p. 5.

[18] S. Francese, R. Bradshaw, and N. Denison, "An update on MALDI mass spectrometry based technology for the analysis of fingermarks - stepping into operational deployment.," *Analyst*, vol. 142, no. 14, pp. 2518–2546, Jul. 2017.

[19] G. Singh, "Determination of Gender Differences from Fingerprints Ridge Density in Two Northern Indian Population of Chandigarh Region," *J. Forensic Res.*, vol. 03, no. 03, pp. 1–3, Mar. 2012.

[20] K. G. Asano, C. K. Bayne, K. M. Horsman, and M. V Buchanan, "Chemical composition of fingerprints for gender determination.," *J. Forensic Sci.*, vol. 47, no. 4, pp. 805–7, Jul. 2002.

[21] M. Nazzaro-Porro, S. Passi, L. Boniforti, and F. Belsito, "Effects of aging on fatty acids in skin surface lipids.," *J. Invest. Dermatol.*, vol. 73, no. 1, pp. 112–7, Jul. 1979.

[22] X. Wang, "Lipid signaling," *Curr. Opin. Plant Biol.*, vol. 7, no. 3, pp. 329–336, Jun. 2004.

[23] K. M. Eyster, "The membrane and lipids as integral participants in signal transduction: lipid signal transduction for the non-lipid biochemist," *Adv. Physiol. Educ.*, vol. 31, no. 1, pp. 5–16, Jan. 2007.

[24] S. A. Saddoughi, P. Song, and B. Ogretmen, "Roles of Bioactive Sphingolipids in Cancer Biology and Therapeutics," in *Lipids in Health and Disease*, Dordrecht: Springer Netherlands, 2008, pp. 413–440.

[25] V. Hinkovska-Galcheva, S. M. VanWay, T. P. Shanley, and R. G. Kunkel, "The role of sphingosine-1-phosphate and ceramide-1-phosphate in calcium homeostasis.," *Curr. Opin. Investig. Drugs*, vol. 9, no. 11, pp. 1192–205, Nov. 2008.

[26] R. R. Watkins, T. L. Lemonovich, and R. A. Salata, "An update on the association of vitamin D deficiency with common infectious diseases," *Can. J. Physiol. Pharmacol.*, vol. 93, no. 5, pp. 363–368, May 2015.

[27] G. J. Schütz, "Molecular Cell Biology," in *TUW*, 4th ed., 2017.

[28] F. Baenke, B. Peck, H. Miess, and A. Schulze, "Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development.," *Dis. Model. Mech.*, vol. 6, no. 6, pp. 1353–63, Nov. 2013.

[29] K. Ishimoto *et al.*, "Sterol-mediated Regulation of Human Lipin 1 Gene Expression in Hepatoblastoma Cells," *J. Biol. Chem.*, vol. 284, no. 33, pp. 22195–22205, Aug. 2009.

[30] D. L. Brasaemle, "*Thematic review series: Adipocyte Biology* . The perilipin family of structural lipid droplet proteins: stabilization of lipid droplets and control of lipolysis," *J. Lipid Res.*, vol. 48, no. 12, pp. 2547–2559, Dec. 2007.

[31] A. Svendsen, "Lipase protein engineering," *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, vol. 1543, no. 2, pp. 223–238, Dec. 2000.

[32] R. C. Davis *et al.*, "Assignment of human pancreatic lipase gene (PNLIP) to chromosome 10q24-q26.," *Genomics*, vol. 11, no. 4, pp. 1164–6, Dec. 1991.

[33] HUGO Gene Nomenclature Committee, "LIPC Symbol Report | HUGO Gene Nomenclature Committee." [Online]. Available: https://www.genenames.org/cgi-

bin/gene_symbol_report?q=data/hgnc_data.php. [Accessed: 30-Sep-2018].

[34] T. G. Kirchgessner, K. L. Svenson, A. J. Lusis, and M. C. Schotz, “The sequence of cDNA encoding lipoprotein lipase. A member of a lipase gene family.,” *J. Biol. Chem.*, vol. 262, no. 18, pp. 8463–6, Jun. 1987.

[35] B. Hube, F. Stehr, M. Bossenz, A. Mazur, M. Kretschmar, and W. Schäfer, “Secreted lipases of Candida albicans : cloning, characterisation and expression analysis of a new gene family with at least ten members,” *Arch. Microbiol.*, vol. 174, no. 5, pp. 362–374, Nov. 2000.

[36] N. Nicolaides, “Skin lipids: their biochemical uniqueness.,” *Science*, vol. 186, no. 4158, pp. 19–26, Oct. 1974.

[37] L. M. Milstone, “Epidermal desquamation,” *J. Dermatol. Sci.*, vol. 36, no. 3, pp. 131–140, Dec. 2004.

[38] B. Emerson, J. Gidden, J. O. Lay, and B. Durham, “Laser Desorption/Ionization Time-of-Flight Mass Spectrometry of Triacylglycerols and Other Components in Fingermark Samples*,” *J. Forensic Sci.*, vol. 56, no. 2, pp. 381–389, Mar. 2011.

[39] K. E. Bloch, “Sterol, Structure and Membrane Function,” *Crit. Rev. Biochem.*, vol. 14, no. 1, pp. 47–92, Jan. 1983.

[40] A. Pappas, “Epidermal surface lipids,” *Dermatoendocrinol.*, vol. 1, no. 2, pp. 72–76, Mar. 2009.

[41] S. Pleik, B. Spengler, T. Schäfer, D. Urbach, S. Luhn, and D. Kirsch, “Fatty Acid Structure and Degradation Analysis in Fingerprint Residues,” *J. Am. Soc. Mass Spectrom.*, vol. 27, no. 9, pp. 1565–1574, Sep. 2016.

[42] P. Hinners, K. C. O’Neill, and Y. J. Lee, “Revealing Individual Lifestyles through Mass Spectrometry Imaging of Chemical Compounds in Fingerprints,” *Sci. Rep.*, vol. 8, no. 1, p. 5149, Dec. 2018.

[43] P. Hinners, K. C. O’Neill, and Y. J. Lee, “Revealing Individual Lifestyles through Mass Spectrometry Imaging of Chemical Compounds in Fingerprints,” *Sci. Rep.*, vol. 8, no. 1, p. 5149, Dec. 2018.

[44] E. Martínez-Pinilla, A. Oñatibia-Astibia, and R. Franco, “The relevance of theobromine for the beneficial effects of cocoa consumption.,” *Front. Pharmacol.*, vol. 6, p. 30, 2015.

[45] “Health and Drugs - Disease, Prescription and Medication by Nicolae Sfetcu (eBook),” *GNU Free Documentation License*, 2014. .

[46] M. J. Arnaud, “Pharmacokinetics and Metabolism of Natural Methylxanthines in Animal and Man,” in *Handbook of experimental pharmacology*, no. 200, 2011, pp. 33–91.

[47] F. Agyemang-Yeboah, H. Asare-Anane, and S. Y. Oppong, *3. Caffeine: The wonder compound, chemistry and properties*, vol. 37, no. 2. .

[48] National Center for Biotechnology Information. PubChem Compound Database, “Caffeine.” [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/caffeine#section=Top. [Accessed: 01-Jan-2019].

[49] D. Laurent *et al.*, “Effects of Caffeine on Muscle Glycogen Utilization and the Neuroendocrine Axis during Exercise [1],” *J. Clin. Endocrinol. Metab.*, vol. 85, no. 6, pp. 2170–2175, Jun. 2000.

[50] A. P. Winston, E. Hardwick, and N. Jaberi, “Neuropsychiatric effects of caffeine,” *Adv.*

Psychiatr. Treat., vol. 11, no. 6, pp. 432–439, Nov. 2005.

[51] P. Holmgren, L. Nordén-Pettersson, and J. Ahlner, “Caffeine fatalities—four case reports,” *Forensic Sci. Int.*, vol. 139, no. 1, pp. 71–73, Jan. 2004.

[52] N. Rodopoulos, O. Wisen, and A. Norman, “Caffeine metabolism in patients with chronic liver disease,” *Scand. J. Clin. Lab. Invest.*, vol. 55, no. 3, pp. 229–242, Jan. 1995.

[53] C. Langbauer and B. angestrebter akademischer Grad, “"Time-course measurements of caffeine and its primary metabolites extracted from fingertips after coffee intake" verfasst von.”

[54] A. Zulli *et al.*, “Caffeine and cardiovascular diseases: critical review of current research,” *Eur. J. Nutr.*, vol. 55, no. 4, pp. 1331–1343, Jun. 2016.

[55] C. Mclean, T. E. Graham, P. Centre, and J. T. Powell Bldg, “Effects of exercise and thermal stress on caffeine pharmacokinetics in men and eumenorrheic women Downloaded from,” *J Appl Physiol*, vol. 93, pp. 1471–1478, 2002.

[56] M. Okuro, N. Fujiki, N. Kotorii, Y. Ishimaru, P. Sokoloff, and S. Nishino, “Effects of paraxanthine and caffeine on sleep, locomotor activity, and body temperature in orexin/ataxin-3 transgenic narcoleptic mice.,” *Sleep*, vol. 33, no. 7, pp. 930–42, Jul. 2010.

[57] E. Martínez-Pinilla, A. Oñatibia-Astibia, and R. Franco, “The relevance of theobromine for the beneficial effects of cocoa consumption.,” *Front. Pharmacol.*, vol. 6, p. 30, 2015.

[58] M. J. Baggott *et al.*, “Psychopharmacology of theobromine in healthy volunteers.,” *Psychopharmacology (Berl).*, vol. 228, no. 1, pp. 109–18, Jul. 2013.

[59] T. N. Jilani and S. Sharma, *Theophylline*. StatPearls Publishing, 2018.

[60] “Theophylline - FDA prescribing information, side effects and uses.” [Online]. Available: https://www.drugs.com/pro/theophylline.html. [Accessed: 27-Nov-2018].

[61] L. M. Juliano and R. R. Griffiths, “A critical review of caffeine withdrawal: empirical validation of symptoms and signs, incidence, severity, and associated features,” *Psychopharmacology (Berl).*, vol. 176, no. 1, pp. 1–29, Oct. 2004.

[62] A. Neuberger and L. L. M. van. Deenen, *Modern physical methods in biochemistry*. Elsevier, 1985.

[63] K. NAGY and K. VÉKEY, “Separation methods,” *Med. Appl. Mass Spectrom.*, pp. 61–92, Jan. 2008.

[64] “mass-to-charge ratioin mass spectrometry, mz,” in *IUPAC Compendium of Chemical Terminology*, Research Triagle Park, NC: IUPAC.

[65] K. Brown, K. H. Dingley, and K. W. Turteltaub, “Accelerator Mass Spectrometry for Biomedical Research,” *Methods Enzymol.*, vol. 402, pp. 423–443, Jan. 2005.

[66] E. Stauffer, J. A. Dolan, R. Newman, E. Stauffer, J. A. Dolan, and R. Newman, “Gas Chromatography and Gas Chromatography—Mass Spectrometry,” *Fire Debris Anal.*, pp. 235–293, Jan. 2008.

[67] J. J. Pitt, “Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry.,” *Clin. Biochem. Rev.*, vol. 30, no. 1, pp. 19–34, Feb. 2009.

[68] A. Bazilio and J. Weinrich, “The Easy Guide to: Inductively Coupled Plasma-Mass

Spectrometry (ICP-MS)," 2012.

[69] R. Wolstenholme, R. Bradshaw, M. R. Clench, and S. Francese, "Study of latent fingermarks by matrix-assisted laser desorption/ionisation mass spectrometry imaging of endogenous lipids," *Rapid Commun. Mass Spectrom.*, vol. 23, no. 19, pp. 3031–3039, Oct. 2009.

[70] S. K. Al-Tarawneh and S. Bencharit, "Applications of Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight (SELDI-TOF) Mass Spectrometry in Defining Salivary Proteomic Profiles," 2009.

[71] "Food Reviews International Overview of the Applications of Tandem Mass Spectrometry (MS/MS) in Food Analysis of Nutritionally Harmful Compounds Stamatia I. Kotretsou & Aglaia Koutsodimou," 2007.

[72] A. S. Ptolemy, E. Tzioumis, A. Thomke, S. Rifai, and M. Kellogg, "Quantification of theobromine and caffeine in saliva, plasma and urine via liquid chromatography–tandem mass spectrometry: A single analytical protocol applicable to cocoa intervention studies," *J. Chromatogr. B*, vol. 878, no. 3–4, pp. 409–416, Feb. 2010.

[73] "Ion-molecule reactions involving methyl isocyanide and methyl cyanide," 2007.

[74] NORMAN Association and Dr. Tobias Schulze, "MassBank | European MassBank (NORMAN MassBank) Mass Spectral DataBase," 2006. .

[75] Virtual Mass Spectrometery Labrotary, "Fragmentation Patterns," 2003. [Online]. Available: http://svmsl.chem.cmu.edu/vmsl/Caffeine/caffeine_fragment.htm. [Accessed: 27-Oct-2018].

[76] D. S. Mantus and G. H. Morrison, "Chemical imaging in biology and medicine using ion microscopy," *Mikrochim. Acta*, vol. 104, no. 1–6, pp. 515–522, Jan. 1991.

[77] T. Alexandrov, "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges," *BMC Bioinforma. 2012 1316*, vol. 13, no. 16, p. S11, Nov. 2012.

[78] L. A. McDonnell, A. van Remoortere, N. de Velde, R. J. M. van Zeijl, and A. M. Deelder, "Imaging Mass Spectrometry Data Reduction: Automated Feature Identification and Extraction," *J. Am. Soc. Mass Spectrom.*, vol. 21, no. 12, pp. 1969–1978, Dec. 2010.

[79] C. A. Prestidge, T. J. Barnes, and W. Skinner, "Time-of-flight secondary-ion mass spectrometry for the surface characterization of solid-state pharmaceuticals," *JPP*, vol. 59, pp. 251–259, 2007.

[80] J. G. Ferwerda, *Charting the quality of forage : measuring and mapping the variation of chemical components in foliage with hyperspectral remote sensing*. [publisher not identified], 2005.

[81] R. M. Elowitz, "Imaging Spectroscopy (Hyperspectral Imaging)." [Online]. Available: http://www.markelowitz.com/Hyperspectral.html. [Accessed: 29-Oct-2018].

[82] C.-I. Chang, *Hyperspectral imaging : techniques for spectral detection and classification*. Kluwer Academic/Plenum Publishers, 2003.

[83] E. A. Jones, S.-O. Deininger, P. C. W. Hogendoorn, A. M. Deelder, and L. A. McDonnell, "Imaging mass spectrometry statistical analysis," *J. Proteomics*, vol. 75, no. 16, pp. 4962–4989, Aug. 2012.

[84] G. J. Edelman, E. Gaston, T. G. van Leeuwen, P. J. Cullen, and M. C. G. Aalders, "Hyperspectral imaging for non-contact analysis of forensic traces," *Forensic Sci. Int.*, vol. 223, no. 1–3, pp. 28–39, Nov. 2012.

[85] G. Elmasry, M. Kamruzzaman, D.-W. Sun, and P. Allen, “Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products: A Review,” *Crit. Rev. Food Sci. Nutr.*, vol. 52, no. 11, pp. 999–1023, Nov. 2012.

[86] T. Adão *et al.*, “Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry,” *Remote Sens.*, vol. 9, no. 11, p. 1110, Oct. 2017.

[87] D.-W. Sun, *Hyperspectral imaging for food quality analysis and control*. Academic, 2010.

[88] D. Harvey, “Analytical chemistery: Overview of the spectroscopic Methods,” San Francisco, California, 94105, USA.

[89] “Introduction to the Electromagnetic Spectrum and Spectroscopy | Analytical Chemistry | PharmaXChange.info.” [Online]. Available: https://pharmaxchange.info/2011/08/introduction-to-the-electromagnetic-spectrum-and-spectroscopy/. [Accessed: 31-Oct-2018].

[90] “Electromagnetic Spectrum - Introduction.” [Online]. Available: https://imagine.gsfc.nasa.gov/science/toolbox/emspectrum1.html. [Accessed: 29-Oct-2018].

[91] G. Bellisola and C. Sorio, “Infrared spectroscopy and microscopy in cancer research and diagnosis.,” *Am. J. Cancer Res.*, vol. 2, no. 1, pp. 1–21, 2012.

[92] “Animation Demonstration No. 1. Interaction of Light with matter When light is incident on a material. - ppt download.” [Online]. Available: https://slideplayer.com/slide/9838582/. [Accessed: 29-Oct-2018].

[93] A. A. Gowen, C. P. O’Donnell, P. J. Cullen, G. Downey, and J. M. Frias, “Hyperspectral imaging – an emerging process analytical tool for food quality and safety control,” *Trends Food Sci. Technol.*, vol. 18, no. 12, pp. 590–598, Dec. 2007.

[94] E. Katz and J. Halámek, *Forensic science : a multidisciplinary approach*. Wiley-VCH Verlag, 2016.

[95] I. T. Jolliffe, *Principal Component Analysis. Second Edition*, vol. 98. 2002.

[96] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987.

[97] L. Spinney, “Nobel Prize controversy | The Scientist Magazine®,” 2002. [Online]. Available: https://www.the-scientist.com/news-analysis/nobel-prize-controversy-52371. [Accessed: 04-Nov-2018].

[98] R. M. Caprioli, T. B. Farmer, and J. Gile, “Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS,” *Anal. Chem.*, vol. 69, no. 23, pp. 4751–4760, 1997.

[99] F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait, “Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers,” *Anal. Chem.*, vol. 63, no. 24, p. 1193A–1203A, Dec. 1991.

[100] A. M. Distler and J. Allison, “5-methoxysalicylic acid and spermine: a new matrix for the matrix-assisted laser desorption/ionization mass spectrometry analysis of oligonucleotides,” *J. Am. Soc. Mass Spectrom.*, vol. 12, no. 4, pp. 456–462, Apr. 2001.

[101] P. Seng *et al.*, “Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry,” *Clin. Infect. Dis.*, vol. 49, no. 4, pp. 543–551, Aug. 2009.

[102] T. R. Sandrin, J. E. Goldstein, and S. Schumaker, “MALDI TOF MS profiling of bacteria at the strain level: A review,” *Mass Spectrom. Rev.*, vol. 32, no. 3, pp. 188–217, May 2013.

[103] D. D. Rhoads, H. Wang, J. Karichu, and S. S. Richter, “The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci,” *Diagn. Microbiol. Infect. Dis.*, vol. 86, no. 3, pp. 257–261, Nov. 2016.

[104] N. Zhong, Y. Cui, X. Zhou, T. Li, and J. Han, “Identification of prohibitin 1 as a potential prognostic biomarker in human pancreatic carcinoma using modified aqueous two-phase partition system combined with 2D-MALDI-TOF-TOF-MS/MS,” *Tumor Biol.*, vol. 36, no. 2, pp. 1221–1231, Feb. 2015.

[105] U. Bahr, M. Karas, and F. Hillenkamp, “Analysis of biopolymers by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry,” *Fresenius. J. Anal. Chem.*, vol. 348, no. 12, pp. 783–791, Apr. 1994.

[106] N. Singhal, M. Kumar, P. K. Kanaujia, and J. S. Virdi, “MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis,” *Front. Microbiol.*, vol. 6, p. 791, Aug. 2015.

[107] M. Vestal and P. Juhasz, “Resolution and mass accuracy in matrix-assisted laser desorption ionization- time-of-flight,” *J. Am. Soc. Mass Spectrom.*, vol. 9, no. 9, pp. 892–911, Sep. 1998.

[108] S. Hosseini and S. O. Martinez-Chapa, “Principles and Mechanism of MALDI-ToF-MS Analysis,” 2017, pp. 1–19.

[109] A. Alanio *et al.*, “Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for fast and accurate identification of clinically relevant Aspergillus species,” *Clin. Microbiol. Infect.*, vol. 17, no. 5, pp. 750–755, May 2011.

[110] D. Bailey, E. P. Diamandis, G. Greub, S. M. Poutanen, J. J. Christensen, and M. Kostrzew, “Use of MALDI-TOF for diagnosis of microbial infections.,” *Clin. Chem.*, vol. 59, no. 10, pp. 1435–41, Oct. 2013.

[111] M. Bucknall, K. Y. C. Fung, and M. W. Duncan, “Practical quantitative biomedical applications of MALDI-TOF mass spectrometry,” *J. Am. Soc. Mass Spectrom.*, vol. 13, no. 9, pp. 1015–1027, Sep. 2002.

[112] F. Cobo, “Application of maldi-tof mass spectrometry in clinical virology: a review.,” *Open Virol. J.*, vol. 7, pp. 84–90, 2013.

[113] M. A. Posthumus, P. G. Kistemaker, H. L. C. Meuzelaar, and M. C. Ten Noever de Brauw, “Laser desorption-mass spectrometry of polar nonvolatile bio-organic molecules,” *Anal. Chem.*, vol. 50, no. 7, pp. 985–991, Jun. 1978.

[114] A. Pirkl, J. Soltwisch, F. Draude, and K. Dreisewerd, “Infrared Matrix-Assisted Laser Desorption/Ionization Orthogonal-Time-of-Flight Mass Spectrometry Employing a Cooling Stage and Water Ice As a Matrix,” *Anal. Chem.*, vol. 84, no. 13, pp. 5669–5676, Jul. 2012.

[115] A. Holle, A. Haase, M. Kayser, and J. H. ¨ Ohndorf, “Optimizing UV laser focus profiles for improved MALDI performance MOTIVATION FOR THIS WORK,” *J. MASS Spectrom. J. Mass Spectrom*, vol. 41, pp. 705–716, 2006.

[116] M. Karas, D. Bachmann, and F. Hillenkamp, “Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules,” *Anal. Chem.*, vol. 57, no. 14, pp. 2935–2939, Dec. 1985.

[117] M. Wiegelmann, K. Dreisewerd, and J. Soltwisch, “Influence of the Laser Spot Size, Focal Beam Profile, and Tissue Type on the Lipid Signals Obtained by MALDI-MS Imaging in Oversampling Mode,” *J. Am. Soc. Mass Spectrom.*, vol. 27, no. 12, pp. 1952–1964, Dec. 2016.

[118] C. Menzel, K. Dreisewerd, S. Berkenkamp, and F. Hillenkamp, “The role of the laser pulse duration in infrared matrix-assisted laser desorption/ionization mass spectrometry,” *J. Am. Soc. Mass Spectrom.*, vol. 13, no. 8, pp. 975–984, Aug. 2002.

[119] R. Chen *et al.*, “N-(1-Naphthyl) Ethylenediamine Dinitrate: A New Matrix for Negative Ion MALDI-TOF MS Analysis of Small Molecules,” *J. Am. Soc. Mass Spectrom.*, vol. 23, no. 9, pp. 1454–1460, Sep. 2012.

[120] R. Shroff and A. Svatoš, “Proton Sponge: A Novel and Versatile MALDI Matrix for the Analysis of Metabolites Using Mass Spectrometry,” *Anal. Chem.*, vol. 81, no. 19, pp. 7954–7959, Oct. 2009.

[121] “9-Aminoacridine matrix substance for MALDI-MS, ≥99.5% (HPLC) | Sigma-Aldrich.” .

[122] National Center for Biotechnology Information. PubChem Compound Database;, “Aminacrine.” [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/9-Acridinamine. [Accessed: 01-Jan-2019].

[123] L. Molin, R. Seraglia, F. R. Dani, G. Moneti, and P. Traldi, “The double nature of 1,5-diaminonaphthalene as matrix-assisted laser desorption/ionization matrix: some experimental evidence of the protonation and reduction mechanisms,” *Rapid Commun. Mass Spectrom.*, vol. 25, no. 20, pp. 3091–3096, Oct. 2011.

[124] N. A. Hagan, C. A. Smith, M. D. Antoine, J. S. Lin, A. B. Feldman, and P. A. Demirev, “Enhanced In-Source Fragmentation in MALDI-TOF-MS of Oligonucleotides Using 1,5-Diaminonapthalene,” *J. Am. Soc. Mass Spectrom.*, vol. 23, no. 4, pp. 773–777, Apr. 2012.

[125] S. Bernès, M. R. Pastrana, E. H. Sánchez, and R. G. Pérez, “1,5-Diaminonaphthalene,” *Acta Crystallogr. Sect. E Struct. Reports Online*, vol. 60, no. 1, pp. o45–o47, Jan. 2004.

[126] A. R. Korte and Y. J. Lee, “MALDI-MS analysis and imaging of small molecule metabolites with 1,5-diaminonaphthalene (DAN),” *J. Mass Spectrom.*, vol. 49, no. 8, pp. 737–741, Aug. 2014.

[127] A. M. Seddon, P. Curnow, and P. J. Booth, “Membrane proteins, lipids and detergents: not just a soap opera,” *Biochim. Biophys. Acta - Biomembr.*, vol. 1666, no. 1–2, pp. 105–117, Nov. 2004.

[128] C. R. Zamarreño, S. Lopez, M. Hernaez, I. Del Villar, I. R. Matias, and F. J. Arregui, “Optical Fiber Refractometers based on Indium Tin Oxide Coatings with Response in the Visible Spectral Region,” *Procedia Eng.*, vol. 25, pp. 499–502, 2011.

[129] J. L. Norris *et al.*, “Processing MALDI mass spectra to improve mass spectral direct tissue analysis,” *Int. J. Mass Spectrom.*, vol. 260, no. 2–3, pp. 212–221, Feb. 2007.

[130] A. J. Karttunen, M. Linnolahti, and T. A. Pakkanen, “Structural principles of polyhedral allotropes of phosphorus.,” *Chemphyschem*, vol. 9, no. 17, pp. 2550–8, Dec. 2008.

[131] Bruker Daltonik GmbH, “Ultraflex III User Manual.” Bremen, Germany, p. 54, 2006.

List of Figures

List of Tables