Infrared spectroscopy of blood-based biofluids

Towards populational probing of common phenotypes

Cristina Leonardo

München 2021

Infrared spectroscopy of blood-based biofluids

Towards populational probing of common phenotypes

Cristina Leonardo

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Cristina Leonardo
aus Enna, Italien

München, den 31.05.2021

Erstgutachter: Prof. Dr. Ferenc Krausz
Zweitgutachter: Prof. Dr. Annette Peters
Tag der mündlichen Prüfung: 21.07.2021

Zusammenfassung

Die Fourier-Transformations-Infrarotspektroskopie (FT-IR) ist ein zeit- und kosteneffiziente Technik, das geeignet ist, gleichzeitig die charakteristischen spektralen Fingerabdrücke aller molekularen Bestandteile chemisch komplexer Proben zu erfassen. Jüngste Untersuchungen haben ihr Potenzial an Bioflüssigkeiten für eine schnelle biomedizinische Diagnostik erforscht. Kommerzielle FT-IR-Spektrometer weisen technische Einschränkungen auf, die sich nachteilig auf ihre Detektionsleistung auswirken. Die meisten der bisher veröffentlichten Studien basieren jedoch auf kleinen Fall-Kontroll-Kohorten, die nicht ausreichende statistische Aussagekraft zu bieten haben und nicht repräsentativ für die allgemeine Bevölkerung sind. Darüber hinaus wurde der Einfluss von demographischen Faktoren und Komorbiditäten auf den Infrarot-Fingerabdrücke menschlicher Bioflüssigkeiten nicht genau verstanden. Die neue feldaufgelöste Spektroskopie (FRS) wurde in unseren Labors entwickelt und hier für die Analyse von menschlichem Blutserum und -plasma getestet. Dies nutzt eine wellenformstabile MIR-Pulslaserquelle mit hoher Brillanz und hohem Dynamikbereich. In Kombination mit der elektrooptischen Abtasttechnologie ermöglicht es die Erfassung des kohärenten elektrischen Feldes der angeregten Moleküle direkt im Zeitbereich.

Diese Dissertation konzentriert sich auf den Vergleich der FRS-Spektroskopie mit einem hochmodernen FT-IR-Spektrometer, um das Potenzial molekularer Infrarot-Fingerabdrücke von menschlichen Bioflüssigkeiten zur Erkennung von Krankheiten zu untersuchen. Sowohl die FRS- als auch die FT-IR-Spektroskopie werden hier auf die menschlichen Blutplasmaproben der KORA-FF4-Studie, eine bevölkerungsbasierte Kohorte mit großem Querschnitt, und über eine unabhängige Fall-Kontroll-Studie validiert. Algorithmen des maschinellen Lernens wurden eingesetzt, um die Erkennungseffizienz und die spektralen Signaturen der einzelnen untersuchten Erkrankungen zu bestimmen. Die Auswirkung allgemeiner demografischer Parameter auf die Infrarot-Fingerabdrücke werden zunächst bei gesunden Personen untersucht, um ihren Einfluss auf die Variabilität zwischen Personen zu ermitteln. Insbesondere Alter und Entzündung werden als Hauptfaktoren identifiziert, die die große spektrale Variabilität der Infrarotsignaturen zwischen Personen beeinflussen. Die Abhängigkeit der Detektionseffizienz von der Anzahl der untersuchten Individuen wird bei der Bewertung der einzelnen Erkrankungen berücksichtigt. Insbesondere zeigt die Analyse, dass mindestens 130 Probanden involviert sein sollten, um eine ausreichende statistische Power der Studie zu gewährleisten. Letztendlich wird über die chemische Fraktionierung im Rahmen der spektroskopischen Erkennung von Krebs berichtet, die ein tieferes molekulares Verständnis der Infrarot-Fingerabdrücke von menschlichem Blutserum und Plasma liefert.

Der Vergleich unabhängiger Kohorten zeigt, dass für die Zielbedingungen spezifische Infrarot-Fingerabdrücke auch für unterschiedliche Studien erzielt werden können, was die Robustheit des Ansatzes unterstreicht. Insbesondere die zeitliche Auflösung der FRS-Spektroskopie erlaubt die Unterscheidung zwischen spezifischen und unspezifischen spektralen Signaturen in unterschiedlichen Zeitfenstern. Diese überlappen sich im Frequenzbereich, was einen wichtigen Vorteil gegenüber der FT-IR-Spektroskopie darstellt. Sowohl FT-IR als auch FRS erweisen sich als etwa gleichermaßen vielversprechend für die Erkennung von Personen, die von Lungenkrebs, Diabetes, Bluthochdruck, hohen Blutfetten betroffen sind, sowie von Personen, die einen Schlaganfall oder Herzinfarkt erlitten haben. Dennoch bietet die FRS-Spektroskopie eine höhere Erkennungseffizienz für verschiedene Prädiabetes-Entitäten sowie

für Brust- und Prostatakrebs. Es ist zu erwarten, dass das diagnostische Potenzial der FRS aufgrund der technischen Verbesserungen, die derzeit in unseren Labors entwickelt werden, deutlich zunehmen wird. Generell hat sich der Infrarot-Fingerabdruck von Blutplasma und -serum als zeiteffektive, robuste und kostengünstige Technik zur Erkennung und Überwachung mehrerer Krankheiten erwiesen, sowohl einzeln als auch gleichzeitig, und stellt damit potenziell ein wertvolles Werkzeug für Anwendungen im klinischen Bereich da.

Abstract

Fourier-transform infrared spectroscopy (FT-IR) is a time and cost-effective technique suited to simultaneously record the characteristic infrared spectral fingerprints of all molecular constituents of chemically complex samples. Recent investigations have explored its potential on biofluids for rapid biomedical diagnostics. However, most of the previously published studies are based on small observational case-control cohorts of clinical patients providing low statistical power and not aimed to assess any general population, often focused on specific medical conditions not providing validation with independent cohorts. Furthermore, it was not well understood what is the impact of demographic factors and comorbidities on the infrared fingerprints of human biofluids. Commercial FT-IR spectrometers present technical limitations, detrimental to their analytical power. The new field-resolved spectroscopy (FRS) has been developed in our laboratories and tested here for the screening of human blood serum and plasma. FRS exploits a waveform-stable few-cycle MIR pulsed-laser source with high brilliance and dynamic range. Combined with electro-optic sampling technology, it allows for the detection of the background-free coherent electric field of excited molecules in the time domain.

This dissertation focuses on the comparison of FRS spectroscopy with a state-of-the-art FT-IR spectrometer to address the potential of infrared molecular fingerprinting of human biofluids. Both FRS and FT-IR spectroscopy are here applied on the human blood plasma samples of the KORA-FF4 study, a large cross-sectional population-based cohort, and validated via an independent case-control study. Machine learning algorithms have been used to determine the detection efficiency and the spectral signatures of each medical condition investigated. The effect of demographic parameters on the infrared fingerprints is first evaluated among healthy individuals to address their impact on the between-person variability, addressing age and inflammation as the main factors. The diagnostic potential of FT-IR and FRS spectroscopy is then addressed for the detection of several common medical conditions, for which the dependence of the detection efficiency on the number of individuals investigated is also considered. In particular, about 130 individuals per group are required to guarantee enough statistical power to the analysis. Ultimately, chemical fractionation is reported in the frame of spectroscopic detection of cancer, providing a deeper molecular understanding of the infrared fingerprints of human blood serum and plasma.

The comparison of independent cohorts shows that infrared fingerprints specific to each medical condition can be obtained for very different study settings, highlighting the robustness of the approach. The temporal resolution of FRS allows the distinction between specific and unspecific spectral signatures in different time windows, providing an important advantage over FT-IR spectroscopy where the signature overlap. FT-IR and FRS are found to be equally promising for the detection of individuals affected by lung cancer, diabetes, hypertension, high blood lipids as well as of individuals who had episodes of stroke or heart attack. Nevertheless, FRS provides higher detection efficiencies for different prediabetes entities as well as for breast and prostate cancer. The diagnostic efficiency of FRS is expected to increase significantly due to the technical improvements currently under development. Ultimately, infrared fingerprinting of human blood biofluids is proved to be a robust, time-effective and efficient technique for detecting and monitoring several medical conditions singularly as well as simultaneously, thus potentially providing a valuable tool for applications in clinical settings.

Contents

List of Acronyms

AC - Agglomerative clustering

AGR - Albumin-globulin ratio

AIC - Akaike information criterion

ALB - Albumin (gene)

ATR-FTIR - Attenuated total reflection-Fourier transform-infrared spectroscopy

AUC - Area under the curve

BCa - Breast cancer

BIC - Bayesian information criterion

BMI - Body mass index

BPH - Benign prostatic hyperplasia

BS - Beam splitter

COPD - Chronic obstructive pulmonary disease

CRP - C-reactive protein

CV - cross-validation

DFG - difference frequency generation

EC - Echo correction

EDTA - Ethylenediaminetetraacetic acid

EM - Expectation maximization

EMF - Electric-field-resolved molecular fingerprint

EOS - Electro-optic sampling

FCG - Fasting capillary glycemia

FN - False negative

FP - False positive

FPG - Fasting plasma glucose

FRS - Field-resolved spectroscopy

FT - Fourier-transform

FT-IR - Fourier-transform infrared spectroscopy

GMM - Gaussian mixture model

GT - Global-T-position

HC - Hilbert centering

HDL - High-density lipoprotein

HP - Haptoglobin (gene)

HPLC - High-performance liquid chromatography

HSA - Human serum albumin

HTPF - High temporal pass filter

HWP - half-wave plate

IC - Interference correction

IDFG - intra-pulse difference frequency generation

IDT - interferometric delay tracking system

IFG - Impaired fasting glucose

IGT - Impaired glucose tolerance

IR - Infrared

IRE - Internal reflection element

KM - K-means

KORA-FF4 - Cooperative Health Research in the Region of Augsburg

L4L - Laser4Life

LCa - Non-small cell lung cancer

LDL - Low-density lipoprotein

LGS - lithium gallium sulfide ($LiGaS_2$)

LOD - Limit of detection

LV - Loading vector

M/F - Males-to-females ratio

MIR - Mid-infrared

NIR - Near-infrared

NSP* - Non-symptomatic individuals

NSP/NGT - Non-symptomatic normal glucose tolerance individuals

OGTT - Oral glucose tolerance test

ORM1 - Alpha-1-acid glycoprotein (gene)

PC - Principal component

PCA - Principal component analysis

PCa - Prostate cancer

POC - Point-of-care

QC - Quality control

QCL - Quntum cascad laser

QWP - quarter-wave plate

ROC - Receiver operating characteristic curve

SEIRA - Surface enhanced infrared absorption

SERS - Surface enhanced Raman scattering

SFG - sum frequency generation

SNR - signal-to-noise ratio

SNR - Signal-to-noise ratio

SOP - Standard operating procedure

SPF - short-pass filter

SPP - Surface plasma polatiron

ST - Standardization

STD - Standard deviation

SVD - Singular value decomposition

SVM - Support vector machine

TN - True negative

TP - True positive

WCSS - Within-custer sum of squares

WHO - World Health Organization

y/o - years old

Yb:YAG - ytterbium-doped with yttrium aluminium garnet ($Y_3Al_5O_{12}$)

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

In 1800 Sir William Herschel demonstrates for the first time the existence of the *thermometrical spectrum* of the sun, later identified as infrared (IR) radiation, with the sole help of a prism and a thermometer [1]. From that moment, it took about 80 years for chemical infrared spectroscopy to emerge and another 80 for the commercialization of the first IR spectrometer. In 1957 Perkin-Elmer placed on the market the Infracord double-beam spectrophotometer [2], able to record spectra from 660 up to 4000 cm^{-1} in only 12 minutes, finally making infrared spectroscopy fast and easily accessible. The same year, Norman K. Freeman introduces the potential of the technique for the analysis of biological samples as well as of tissues in his book *"Advances in Biological and Medical Physics"* [3]. Infrared spectroscopy is soon applied for disease detection in humans, leading to a comparable growth rate in the number of publications about this topic as for IR spectroscopy in general (Figure 1.1).

The first paper about IR spectroscopy appears in Science in 1927 [1], but the number of publications grows slowly in the next years. It is only in 1963 that the research around IR spectroscopy starts to flourish. In the same year, the first paper reporting IR spectroscopy on a human biofluid for disease detection is published [4]. In this paper, the IR spectra of urine samples of one patient affected by Hurler's syndrome and three siblings with Morquio-Ullrich´s disease are compared with the spectra of the molecules expected to be up-regulated by the respective conditions: chondroitin sulfate B in both cases and altered hyaluronic acid in the last. For the first time, the potential of IR spectroscopy for the detection of disease-induced biological changes in the chemical composition of human biofluids is reported. This happens already before the advent of microcomputers able to perform Fourier-transform (FT), which lead to an additional boost in the commercialization of the more advanced FT-IR spectrometers based on the Michelson interferometer technology. The first FT-IR spectrometer, by Digilab (Model FTS-14), is available on the market in 1969 [5]. By the year 2000, the performance-to-prize ratio of these devices has been incredibly boosted, making the applicability of FT-IR spectroscopy in clinical settings increasingly interesting.

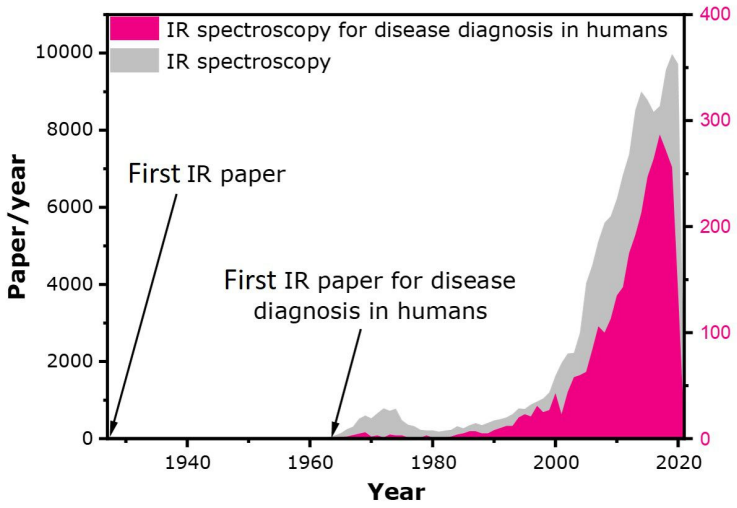

Figure 1.1: Number of publications about infrared spectroscopy in general and applied to disease diagnosis in human biosamples per year. Source: PubMed [6].

In the last twenty years, the potential of FT-IR spectroscopy on biofluids and tissues for the screening of the human health status has been largely explored, particularly for the early detection of specific medical conditions and as a tool for treatment monitoring. Indeed, as a result of the huge advances in medicine, developed nations have reached a higher quality of life leading to longer life expectancy [7]. As a consequence of aging populations, an increase of chronic diseases has been recorded, leading to high demand for new fast and reliable throughput diagnostic tools. Highly sensitive techniques, such as mass spectrometry and nuclear magnetic resonance, have been proposed for the identification of specific biomarkers for several conditions [8–14]. On the other hand, FT-IR spectroscopy can acquire information about concentration and structure of all molecular constituents of the sample under analysis in one single measurement, offering an advantageous approach compared to biomarker-targeted techniques. One of the main advantages of FT-IR is its high potential for the diagnosis of multiple conditions in one fast measurement making it suitable for fast patients triaging and personalized medicine [15, 16]. Moreover, it is simple to operate, reagent-free, label-free, non-destructive and easily adaptable to clinical environments [16]. Several studies have been proven the potential of infrared spectroscopy in different biofluids, such as human blood serum [17] and plasma [18], whole blood [19], sputum [20], bile [21], amniotic fluid [22], cerebrospinal fluid [23], urine [24], saliva [25] and tears [26]. The biofluid determines how invasive and costly IR fingerprinting for disease diagnosis is. Human blood serum and plasma are the most advantageous being easily available and, more importantly, the only ones able to carry information about the health status of every organ in the body.

Many efforts have been done to prove the advantages of the technique, but few in the direction of actual clinical applications. To this end, further improvements are still needed. FT-IR must ultimately be proven to improve patients outcomes and alleviate financial strains on the healthcare system. The cooperation between instrument developers, spectroscopists and clinicians is therefore essential to this end [16]. So far, only a few major programs have been

moving toward applications of IR spectroscopy on large scale clinical trials, such as: *Biotech Resources*, which uses attenuated total reflection-Fourier transform-infrared spectroscopy (ATR-FTIR) of human blood cells for the screening of malaria in low-income areas [27]; *Cireca Theranostics*, which has developed "spectral histopathology" protocols for the analyses of tissue sections as support to histopathologists in the decision-making process [28, 29]; *Glyconics*, which together with *Spectrolytics*, a German spectrometers development company, aim at producing a hand-held point-of-care (POC) FT-IR device specifically for clinical settings for the analysis of chronic obstructive pulmonary disorder (COPD) and diabetes [30–32]; *Clinspec Diagnostics Ltd*, implementing the clinical translation of ATR-FTIR to identify and classify brain cancer from blood serum samples, investigating also its costs and health benefits [33].

Several reasons limit the approval of FT-IR spectroscopy in clinical practice, such as the general lack of standardization and, especially, of validation in multi-center clinical trials [15]. Most of the several studies published to prove the potential of IR spectroscopy for disease detection are proof-of-principle, usually based on small retrospective case-control studies. In these cohorts, the individuals are sampled based on the presence (cases) or absence (controls) of the target condition. On the other hand, in prospective cohorts, people without the "outcome" of interest are sampled in specific conditions (in a population, among the patients of specific clinics, etc.) and a variety of parameters that might be relevant to the development of the target condition are recorded over a specified period; the individuals who develop the condition are the cases and the rest are internal controls [34]. Both samplings are affected by different sources of selection biases leading to cohorts of individuals systematically different from the population they intend to represent. The most important selection biases are: non-response bias, due to the dependence of the sampling from non-responders and responders; volunteer-bias, which considers the potential differences between those who volunteer (reported to be more sociable, from higher social classes and more educated) and the general population; ascertainment-bias, which is the systematic difference between the actual data and the reported one, either due to the study participants (response-bias) or to the investigators (assessment or observer-bias) [35].

Case-control studies are usually retrospective and more cost-effective because the collection is focused only on the individuals of interest, usually among clinic patients, making them specific to a target population [34]. Cross-sectional studies, on the other hand, are usually prospective and individuals are sampled either from national registers or in a random fashion to represent the entire population [35, 36]. Both case-control and cross-sectional studies are referred to as "observational" because the investigator simply observes. Population-based studies are more likely to be representative of the whole population, thereby minimizing selection biases, especially if random sampling is adopted [35]. Prospective cohorts reduce these biases as well because cases and controls are collected independently from the prognosis [15]. Therefore, prospective cross-sectional population-based cohorts are the ideal study to evaluate the infrared fingerprints of common phenotypes in the general population. Moreover, as opposed to case-control studies which focus on one outcome in each study, cross-sectional cohorts examine various parameters allowing for prospective case-control-like analysis [34]. The main drawback of such studies, which is the reason why they are so rarely used, is the long time required for sample collection. This is also due to the need for recruiting more individuals to compensate for eventual losses of subjects in follow-up studies. Moreover, these cohorts are not suited for the analysis of rare conditions that would be underrepresented as only a few individuals will develop them during the study [34, 37].

In the past 20 years, no major technological improvements have been applied to FT-IR spectroscopy. The thermal radiation sources feature quite a low brilliance, with detrimental effects on the signal-to-noise ratio (SNR). Recently, new sources with higher photon flux density have been proposed, such as quantum cascade lasers (QCL) and coherent femtosecond broadband sources [38–41]. Synchrotron infrared sources provide 100-1000 times higher photon flux density and have been shown to provide better disease diagnostic efficiencies in several studies [42–46]. However, their accessibility is very limiting. The high noise and the limited dynamic range of the detectors in commercial FT-IR spectrometers are also limiting. Frequency up-conversion detectors can provide higher quantum-efficiency and sensitivity [5]. Other sources of technical noise, such as interferometer instabilities, contribute to lowering the SNR and are usually compensated by balanced detection, active stabilization or fast data-acquisition [47–51].

To boost the efficiency of commercial IR spectrometers, a new technique has been recently implemented in our research group: *field resolved spectroscopy* (FRS) [52]. A waveform-stable few-cycle MIR pulsed-laser is used as a high brilliance source. Combined with electro-optic sampling (EOS), it allows the detection of the coherent *electric-field-resolved molecular fingerprint* (EMF) of the excited molecules in a background-free fashion. This extensively reduces the influence of the source fluctuations on the signal of interest and allows the use of higher brilliance sources without the drawback of saturation at the detectors. These advantages allow the detection of less abundant molecules in their natural aqueous environment pushing the limit of detection (LOD) down to 200 ng/mL, 40 times lower than state-of-the-art commercial FT-IR devices (8 μg/mL) and 5 orders of magnitude lower than the most abundant molecule, human serum albumin. Moreover, FRS performances have been demonstrated also for the investigation of intact biological systems with high optical and physical thicknesses (FT-IR upper limit is 10μm), such as human THP-1 leukemic-monocyte-like cell line and on goat willow (*Salix caprea*) leaves with 120 μm thickness [52]. It is therefore a very powerful tool for biology, biomedicine, pharmaceutical and ecological applications. Further developments are ongoing, such as the broadening of the spectral coverage, currently between 1000 and 1500 cm^{-1}, and the implementation of fast scanning able to "freeze" the excitation pulse fluctuations.

In this dissertation, FT-IT and FRS spectroscopy are applied to human blood plasma samples collected in the frame of KORA-FF4 (cooperative health research in the region Augsburg), a prospective cross-sectional population-based cohort. The samples have been collected by the Helmholtz Zentrum in 2013-2014 as a continuation of a longitudinal monitoring study of epidemiology, health economics and health care of a German population [53, 54]. The collection has been performed via stratified random sampling in seventeen communities in Augsburg following the German guidelines for *Good Epidemiological Practice* [55]. The participants underwent in-person interviews and medical examinations leading to a platform with more than 10.000 variables, including socio-demographic parameters, risk factors, medical history and medications [54, 56, 57]. The main results obtained for KORA-FF4 via FT-IR are compared with an independent prospective case-control study for cancer diagnosis, Laser 4 Life (L4L), by a consortium between Ludwig-Maximilians-Universität, the Max Planck Institute for Quantum Optics and multiple international clinical centers in Europe. This is based on human blood serum collected in 2018-2019 from clinical patients in the area of Munich. The validation with an independent study is often missing in recent publications on this topic [58]. The large preanalytical error, amounting to about 60% of the total errors in clinical practice [59],

is a potential obstacle in the comparison of independent cohorts. However, this is shown to be negligible compared to the biological diversity between individuals, which makes IR spectroscopy suitable for screening purposes [60].

Overall, this dissertation presents an investigation of the IR signatures of demographics and other common factors in human blood biofluids of healthy individuals and the influence of these parameters on the diagnosis of common medical conditions. To the best of our knowledge, this analysis has never been carried out to this extent on a population-based cohort before.

1.1 Outline of the dissertation

The tools used in this dissertation are introduced in *chapter 2*: from common FT-IR spectrometers and other vibrational spectroscopy techniques to the description of the FRS set-up, with respective advantages and disadvantages. Ultimately, the machine learning algorithms used for the analysis of both FT-IR and FRS data are discussed.

In *chapter 3* the technical and biological noise are discussed for both techniques. The specificity of the IR analysis of human blood biofluids has been questioned because of the large between-person variability due to demographics and other common parameters (age, gender, smoking status, etc.). In particular, inflammation has been addressed as one of the factors with the strongest impact on the spectral signature of proteins [61]. Diseases are normally associated with molecular concentrations varying beyond the normal physiological levels. However, the large biological variability between individuals can mask their signatures, especially at the earliest stages. *Chapter 3* addresses the role of these parameters on the between-person spectral variability among the healthiest individuals of KORA-FF4 and L4L cohorts.

Chapter 4 focuses on the detection of endpoint and intermediate conditions. Prospective randomized studies are often thought to eliminate the influence of confounding variables: since each type of exposure is assigned by chance, they should be equally distributed among cases and controls [34]. However, the comparison of the fingerprints of the same diseases in KORA-FF4 and L4L, which are very different in nature, shows that if cases and controls are not matched other factors might affect the fingerprints [15]. Matching is shown to isolate phenotype-specific signatures, necessary to address the actual diagnostic efficiency of different diseases.

Ultimately, the high chemical complexity of human blood biofluids is associated with a large dynamic range of concentrations of the different biomolecules. The strong signals of the most abundant molecules cover the spectral signatures from the low abundant ones [16, 61]. To solve this issue, chemical fractionation is applied in *chapter 5* on human blood serum samples collected in a pilot case-control study for cancer detection [62]. This allows a deeper understanding of the IR fingerprints at a molecular level.

Chapter 2

Infrared spectroscopy and machine learning algorithms

The techniques and machine learning algorithms used in this dissertation are introduced in this chapter. First, a theoretical description of molecular vibrations and the related spectroscopic transitions is given. This is followed by a short description of the technical aspects of commercial FT-IR spectrometers, with related advantages and limitations. Similar techniques increasingly used for the analysis of human biofluids are briefly introduced. To conclude the list of vibrational spectroscopic techniques, FRS is ultimately discussed. In particular, a description of the technology behind is proposed and advantages and disadvantages of the technique are reported. Ultimately, the machine learning and the matching algorithms used in the following chapters are introduced.

2.1 Quantum mechanics definition of vibrational modes and excitations

This section aims at introducing the theoretical aspects of vibrational spectroscopy, starting from the quantum mechanic definition of normal modes and concluding with the related spectroscopic transitions.

2.1.1 Definition of molecular normal modes

Before going into the technical aspects of vibrational spectroscopy, a theoretical description of molecular vibrations is presented, from the description of a free particle to the harmonic oscillator. The following discussion refers to the more thorough explanation of this topic in [63] and [64].

To start with, it is essential to introduce the most relevant basic notion of quantum mechanics: all the properties of a system depending on space (**r**) and time (t) are described by a *wavefunction*. A given wavefunction $\Psi(\mathbf{r}, t)$ is a complex mathematical function which, if squared, returns the *probability density*, namely the probability to find the system in a given unit of volume. The total probability is therefore:

$$\int P(\mathbf{r}, t) d\sigma = \int \Psi^*(\mathbf{r}, t)\Psi(\mathbf{r}, t) d\sigma = \langle \Psi | \Psi \rangle \tag{2.1}$$

where Ψ^* is the complex conjugate of Ψ, respectively denoted as $\langle \Psi |$ and $| \Psi \rangle$ in bra-ket notation.

$P(\mathbf{r}, t)$ must have a single value for each position. Therefore, the wavefunction needs to be single-valued as well and its integral must always exist and be finite. Moreover, it must be continuous in space, with continuous first derivative except at the system boundaries. Another essential concept of quantum mechanics is that any measurable physical property A is the *expectation value* of a correspondent Hermitian *operator* $\hat{A}$, expressed as:

$$A = \langle \Psi | \hat{A} | \Psi \rangle \tag{2.2}$$

where Hermitian means that, given two wavefunctions Ψ_a and Ψ_b, it follows that $\langle \Psi_b | \hat{A} | \Psi_a \rangle = \langle \Psi_a | \hat{A} | \Psi_b \rangle^*$. The quantum mechanical operator associated to the direction $\mathbf{r}$ is simply $\hat{\mathbf{r}} = \mathbf{r}$, while the one associated to the momentum $\mathbf{p}$ is $\hat{\mathbf{p}} = -i\hbar\hat{\nabla}$, where $\hbar$ is the Planck' constant h divided by 2π and $\hat{\nabla} = (\partial/\partial x + \partial/\partial y + \partial/\partial z)$ is the gradient operator. One of the most important operators to describe any system is the *Hamiltonian*, which returns the total energy E as the sum of the kinetic (T) and potential energy (V). In particular, from the definition of $\hat{\mathbf{p}}$, it follows that the operator kinetic energy is:

$$\hat{T} = \frac{\hat{\mathbf{p}}^2}{2m} = \Big(-\frac{\hbar^2}{2m} \Big) \hat{\nabla}^2 \tag{2.3}$$

where $\hat{\nabla}^2 = (\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)$ is the *Laplacian* operator.

In 1926, Erwin Schrödinger finds an elegant solution to derive the wavefunction of a system, the *time-dependent Schrödinger equation*:

$$\hat{H}\Psi = i\hbar \frac{\partial \Psi}{\partial t} \tag{2.4}$$

which, from the definition of Hamiltonian and from equation 2.3, can be reformulated as:

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \hat{\nabla}^2 \Psi + V(\mathbf{r})\psi \tag{2.5}$$

which has the following solution:

$$\Psi = \psi(\mathbf{r}) exp(-2\pi i E t/\hbar) \tag{2.6}$$

The associated probability, described in equation 2.1, only depends on the radial part of the solution and is equal to $|\psi(\mathbf{r})|^2$. If the Hamiltonian only depends on the position, namely $\hat{H} = \hat{H}_0(\mathbf{r})$, the equation 2.5 becomes:

$$\hat{H}_0 \psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{2.7}$$

which is the *time independent Schrödinger equation*. The solutions of this *eigenvalue equation* are *eigenfunctions* $\psi_k(\mathbf{r})$, which for an electron in an atom or molecule are the set of orbitals and represent an *eigenstate* associated with a specific *eigenvalue* E_k of the system. The Hamiltonian operator returns a complete set of *orthogonal* functions ($\langle \psi_i | \psi_k \rangle = 0$) which can be multiplied by a nonzero value to be *normalized* ($\langle \psi_i | \psi_i \rangle = 1$) giving *orthonormal eigenfunctions*. Any linear combination of eigenfunctions, called *superposition of states*, is also an eigenfunction.

The time-independent Schrödinger equation of a free particle in a constant potential, which can be set to zero, is given by equation 2.7 substituting E with the kinetic energy as expressed

in equation 2.3 and returns not-quantized energies. However, if the particle is confined in a box of length l with potential energy zero inside the box and infinite outside, the wavefunction and the corresponding energy are quantized according to the first *quantum number n*, which can have only integer values. Both the number of nodes and the uniformity of the probability distribution increase with n. Moreover, the superposition of states can give wavefunctions more localized in space, called *wavepackets*. The probability $|\psi_n|^2$ is zero outside the box for infinitely high walls. However, if the height of the box is finite, the probability of finding the particle outside the box is low but not zero because of the purely quantistic *tunnel effect*.

The case of an electron in a chemical bond resembles the one of a particle in a box with finite walls. The potential can be approximated to the one of a harmonic oscillator (e.g. a spring connecting two masses), which in one dimension is $V(x) = \frac{1}{2}kx^2$, with x the displacement from the equilibrium mean length and k the force constant. An harmonic oscillator of reduced mass $m_r = m_1 m_2/(m_1 + m_2)$ has an oscillation frequency $\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m_r}}$. Using the defined $V(x)$, the one-dimensional Hamiltonian returns the following eigenvalues:

$$E_n = \left(n + \frac{1}{2}\right)h\nu \tag{2.8}$$

The minimum energy, namely the *zero-point energy*, is therefore $(1/2)\hbar\omega$. As for the case of a particle in a box, also for the harmonic oscillator eigenvalues and wavefunctions are quantized according to n, which determines the number of nodes and the spatial distribution of probabilities that, as before, becomes more uniform for increasing values of n. Compared to the case of a particle in a box, the dependence of E from n is linear and the wavefunctions have a more complex shape. As described before, a wavepacket is a linear combination of harmonic-oscillator wavefunctions. Because of the temporal dependence $exp(-2\pi i E_n t/\hbar)$ in the complete wavefunction $\chi_n(x)$ (see equation 2.6), a wavepacket oscillates in the potential well with the classical frequency ν.

In a molecular system, each vibration is the complex collective motion of many nuclei together, but their analysis can be simplified assuming that the vibrational potential is a harmonic function of the atomic coordinates. Therefore, a set of *normal coordinates* ξ, given by the linear combination of the atoms' coordinates, can disentangle the vibrational modes (or *normal modes*) along single normal coordinates. In this case, the potential can be written as $V = \frac{1}{2}\Sigma_i k_i \xi_i^2$. The wavefunctions of an harmonic oscillator are:

$$\chi_n(x) = N_n H_n(u) exp\left(-\frac{u^2}{2}\right) \tag{2.9}$$

with N_n a normalization factor, H_n an *Hermite polynomial* and u the dimensionless positional coordinate given by:

$$u = \frac{x}{\sqrt{\hbar/2\pi m_r \nu}} \tag{2.10}$$

A molecule with N atoms has $3N$ degrees of freedom, 6 of which are rotations and $3N - 6$ of which are vibrational modes ($3N - 5$ for linear molecules with only two rotational degrees of freedom). In the harmonic approximation, the vibrational wavefunction of each of the $3N - 6$ harmonic oscillators are:

$$X(u_1, u_2, ...) = \Pi_{i=1}^{3N-6} \chi_{n(i)}(u_i) \tag{2.11}$$

with corresponding energies:

$$E_v ib = \Sigma_{i=1}^{3N-6}\left(n_i + \frac{1}{2}h\nu_i\right) \tag{2.12}$$

with n_i the excitation level corresponding to the normal mode i. Despite the harmonic oscillator nicely describes the normal modes of molecules, a better approximation is given by the *Morse potential* [65]:

$$V_M(\xi) = D_e\left(1 - exp\left(-\sqrt{\frac{k_e}{2D_e}}(\xi - \xi_e)\right)\right)^2 \tag{2.13}$$

with D_e the dissociation potential calculated from the zero-point energy, ξ_e the coordinate at equilibrium and k_e the force constant at this position. This potential takes into account the anharmonicity of molecular vibrations which originates from the steeper increase of the potential energy when the system goes to too small inter-atomic distances and to the possibility to stretch the system more strongly than a harmonic oscillator reaching longer inter-atomic distances. One of the major differences with the harmonic oscillator approximation, where the energy gaps are always $\Delta E = h\nu$ (from equation 2.8), is that here the gaps are increasingly smaller at increasing n:

$$\Delta E = E_{n+1} - E_n = h\nu_0 - (n+1)\frac{h\nu_0^2}{2D_e} \tag{2.14}$$

2.1.2 Spectroscopic transitions between normal modes

The harmonic and anharmonic oscillator approximations describe *stationary states* via Hamiltonians which do not depend on time. To describe the transition between vibrational states, stimulated by light in the IR spectral range (200 - 5000 cm^{-1}), the time-dependent Schrödinger equation 2.5 needs to be considered. In particular, the Hamiltonian will be the sum of two contributions:

$$\hat{H} = \hat{H}_0 + \hat{H}'(t) \tag{2.15}$$

with $\hat{H}_0$ the Hamiltonian of the unperturbed system and $\hat{H}'(t)$ describing the perturbation due to the incident electric field of light. In the simple case of a diatomic molecule with vibrational coordinate x, the perturbation term of the Hamiltonian can be written as:

$$\hat{H}'(x,t) \approx -\mathbf{E}(t)\cdot\boldsymbol{\mu} \tag{2.16}$$

where $\mathbf{E}(t)$ is the electric field of the IR radiation and $\boldsymbol{\mu}$ is the molecular dipole moment. The strength of the transition between an initial level n and a final level m is $|\boldsymbol{E_0}\cdot\boldsymbol{\mu}_{mn}|^2$, with $\boldsymbol{E_0}$ the amplitude of the electric field and $\boldsymbol{\mu}_{mn} = \langle\chi_m|\hat{\mu}|\chi_n\rangle$, the matrix element of the electric dipole operator. The amplitude of the molecular electric dipole can be written in a Taylor series as:

$$|\boldsymbol{\mu}(x)| = |\boldsymbol{\mu}\mathbf{0}| + (\partial|\boldsymbol{\mu}|/\partial x)_0 x + \frac{1}{2}(\partial^2|\boldsymbol{\mu}|/\partial x^2)_0 x^2 + ... \tag{2.17}$$

from which the transition dipole moment $\boldsymbol{\mu}_{mn}$ can be re-formulated. In a poliatomic molecule with $3N-6$ normal modes it will have the form:

$$\boldsymbol{\mu}_{mn} = \Sigma_{i=1}^{3N-6}\left(\frac{\partial\boldsymbol{\mu}}{\partial x_i}\right)_0\langle\chi_{m(i)}|x_i|\chi_{n(i)}\rangle + \frac{1}{2}\Sigma_{i=1}^{3N-6}\Sigma_{j=1}^{3N-6}\left(\frac{\partial^2\boldsymbol{\mu}}{\partial x_i\partial x_j}\right)_0\langle\chi_{m(i)}\chi_{m(j)}|x_i x_j|\chi_{n(i)}\chi_{n(j)}\rangle + ... \tag{2.18}$$

The first term $(\partial\boldsymbol{\mu}/\partial x_i)_0\langle\chi_{m(i)}|x_i|\chi_{n(i)}\rangle$ accounts for the main vibrational absorption bands at respective energies $h\nu$, while the second terms accounts for the weaker *overtone* transitions at $2h\nu$. In particular, from the first term it is evident that vibrational absorption can happen only if both parts in the first term are nonzero, which for the matrix element happens only for $m_i - n_i = \pm 1$. Therefore, the selection rules for vibrational spectroscopy are:

$$\left(\frac{\partial|\boldsymbol{\mu}|}{\partial x}\right)_0 \neq 0 \tag{2.19}$$

$$m - n = \pm 1 \tag{2.20}$$

In particular, from equation 2.19 it is evident that IR absorption is prohibited for transition totally symmetric to the molecular structure because they do not induce a change in the electric dipole of the molecule.

The possibility to disentangle vibrations from small functional groups or atoms largely simplifies the interpretation of infrared spectra of macromolecules (Table 2.1), especially for complex samples such as biofluids (Figure 2.1). For example, in proteins the vibrations of the amide group can be distinguished from the side chains and originates four main bands due to:

1. $N - H$ stretching mode (Amide A, 3280 - 3300 cm^{-1});

2. $C = O$ stretching mode with contributions from in-phase bending of $N - H$ bond (Amide I band, 1650 - 1660 cm^{-1} in α-helical structures and 1620 - 1640 cm^{-1} in β-sheets);

3. $N - H$ in-plane vibrations coupled with $C - N$ and $C - C$ stretching and $C = O$ in-plane bending (Amide II band, 1540 - 1550 cm^{-1} in α-helical structures and 1520 - 1525 cm^{-1} in β-sheets);

4. $N - H$ in-plane bending coupled with $C - N$ stretching as well as $C - H$ and $N - H$ deformation (Amide III band, 1200 - 1350 cm^{-1}).

The interactions with other peptide groups, as well as with solvent molecules such as water, create exciton-like couplings with the consequent splitting of the absorption bands. Despite the contributions to the absorption spectra are more complex, the break-down of different bands as the collection of vibrations from a few atoms is not too far from the actual contributions and it allows to attribute the absorption bands to specific functional groups. The main biomolecules in biological samples are built on specific building-blocks based on different functional groups, such as: amino acids, based on the amino group ($-HN - C - CO-$); lipids, such as fatty acids, phospholipids, glycerolipids, sterols (such as cholesterol) and others, all based on saturated and unsaturated hydrocarbon structures ($C - C$, $C = C$ and $C - H$); carbohydrates, based on carbon, oxygen and hydrogen atoms ($C - C$, $C - H$, $C - OH$ and $C - O - C$); nucleotides, with specific structures and interactions in DNA and RNA molecules, based on a pentose sugar, a nitrogenous base and a phosphate group, each with specific vibrational signatures. In particular, in the absorption spectra of human blood serum and plasma, the vibrations can be mainly attributed to specific biomolecules based on their structure and abundance (Table 2.1 and Figure 2.1).

Band (cm^{-1})	Molecular vibration	Main biomolecules	Color in Figure 2.1
1120 - 1170	$\nu(C=O)$, $\nu(C-O-C)$	*carbohydrates*	green
1200 - 1400	$\nu(C-N)$ - Amide III	*proteins*	purple
1240	$\nu_{as}(P=O)$	*nucleic acids*	orange
1400	$\nu(COO^-)$ - Amide III	*amino acids*	light purple
1500 - 1600	$\delta(N-H)$ - Amide II	*proteins*	purple
1600 - 1700	$\nu(C=O)$ - Amide I	*proteins*	purple
1730 - 1760	$\nu(C=O)$	*fatty acids*	yellow
2840 - 2860	$\nu_s(CH_2)$	*lipids*	light yellow
2865 - 2880	$\nu_s(CH_3)$	*lipids*	light yellow

Table 2.1: Association of molecular vibrations in aqueous solutions with the main classes of biomolecules present in human blood serum and plasma, as highlighted in Figure 2.1. The Amide bands arise from the coupling of more vibrational modes; here, only the main ones are highlighted for simplicity. Symbols: ν - stretching, ν_s - symmetric stretching, ν_{as} - asymmetric stretching, δ bending.

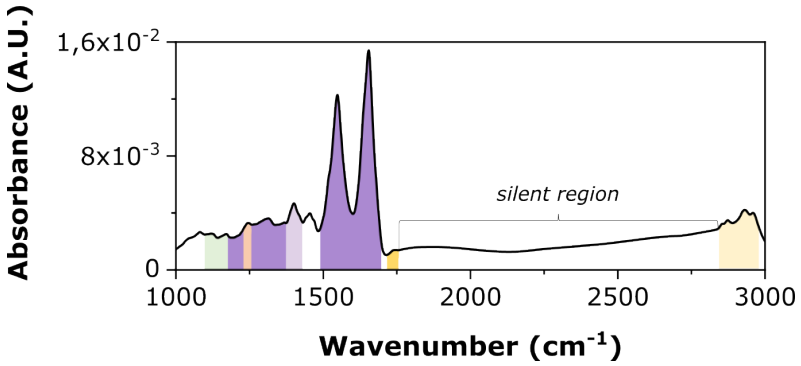

Figure 2.1: Example of the absorption spectrum of a human serum sample highlighting the main biomolecules contributing in the corresponding spectral region, as reported in Table 2.1. No major absorption bands arise in the silent region between 1800 and 2750 cm^{-1}.

2.2 Infrared spectroscopy

This section introduces the basic concepts of IR-based spectroscopic techniques. In particular, the most common techniques employed for the analysis of human biofluids for bio-medical investigations are described, with a particular focus on the two techniques used in this dissertation: FT-IR and FRS spectroscopy.

2.2.1 Technology of commercial spectrometers

In this section, the technical aspects of the most common IR spectrometers commercially available are illustrated with relative advantages and disadvantages.

Every spectrometer has three fundamental components: a source that emits homogeneous radiation in the spectral range of interest, a method/device able to resolve the different wavelets of light and a detector to record the light intensity. The typical sources of IR spectrometers are

thermal radiation sources, solid materials heated to turn incandescent and emit radiation with an energy distribution similar to the Plank' distribution law for the black-body [5]:

$$E_\lambda d\lambda = \frac{c_1 \lambda^{-5}}{e^{(c_2/\lambda T)} - 1} d\lambda \tag{2.21}$$

where $c_1 = 2\pi c^2 h$ and $c_2 = ch/k$; c is the speed of light, h is the Plank's constant and k the Boltzmann's constant. Among the most common sources, there are the Nernst Glower, based on rare earth oxides, and the Globar, a silicon carbide rod. The central frequency λ_c is inversely proportional to the temperature, which is usually set to deliver lower λ_c to guarantee a more homogeneous energy distribution at all wavelengths [66]. However, this is detrimental to the total intensity and the source photon-density flux, the *brilliance*.

The existence of infrared radiation of the solar spectrum has been proven by Sir William Herschel in 1800 via a simple thermometer and a prism. In 1880, Samuel Pierpont Langley introduced a more sophisticated detection system, the bolometers, with which he performed the first accurate measurement of IR wavelengths of the solar spectrum [1]. Nowadays, there are two types of IR detectors used in commercial spectrometers: *thermal detectors*, which measure the heating with the same performance at all frequencies, and *photodetectors*, sensitive to the radiation intensity. Examples of thermal detectors are: thermocouples, which exploit the temperature-dependent voltage of the junction between two dissimilar metals; bolometers, based on a material with a resistance that changes with the temperature; pyroelectric detectors, which uses pyroelectric materials with temperature-dependent polarization which, placed between two electrodes, act as capacitors and convert the thermal changes in voltage; pneumatic detectors, such as the Golay cell, based on a gas-filled chamber where an absorbing material heats the gas upon IR absorption increasing the pressure and deforming a membrane which modulates the incident light of a photodiode. The lasts have a higher sensitivity but are also more expensive. Among the most used photodetectors are photoconductive cells, where the detector material is a semiconductor with a narrow-band gap which features an increase in the electrical conductivity upon IR radiation absorption. Response time and sensitivity of photodetectors can be much higher than thermal ones, but they usually need to be cooled to reduce the thermal noise. Moreover, they are not suited for the detection of long wavelengths with too low energy for electron excitation.

Another fundamental part of any spectrometer is a device able to resolve different wavelengths, the most common being dispersive optics such as prisms and gratings. Both split the different wavelengths of broadband spectra in different directions via dispersion and diffraction effects allowing their isolated detection via the use of slits. The simple sources and detectors already available, combined with the well-known technology of dispersive optics, have made possible the commercialization of the first dispersive spectrometer already in the 1940s, spreading IR spectroscopy in laboratories. The first relatively cheap spectrometer on the market was the Infracord by Perkin-Elmer (1957), with spectral coverage from 600 to 4000 cm^{-1}, where the lower limit is due to the optical properties of the sodium chloride prism, later replaced by potassium bromide (400 cm^{-1}) or cesium iodide (200 cm^{-1}). The use of grating has allowed reaching even lower wavenumbers.

However, dispersive optics have major limitations and have therefore been replaced by the *Michelson interferometer*. This is based on a beam splitter (BS) that divides the beam from the source in two halves (Figure 2.2): one goes to a fixed mirror (M1) and is sent back to the

source and the detector via the same beam splitter; the other half goes to a mirror (M2) placed on a movable stage (Δd) and is also sent back to the source and the detector via the beam splitter. The two arms will therefore give an interference pattern that can be converted in a spectrum via Fourier transform. In particular, since the electric field of light of the two harms are propagating with the same orientation, their interference can be described as the sum of their scalar values. Each electric field can be expressed as $E_i(z,t) = E_{out}e^{i(\omega t - \kappa z - 2\kappa d_i)}$, with $i = 1, 2$ for the beam reflected at M1 and M2 respectively and E_{out} the field of each beam at the interferometer output, which takes into account the respective phase due to the fact that, for $\Delta d = 0$, each beam travels the distance d_i twice. Knowing that the wavelength is defined as $\lambda_i = \frac{2\pi}{|\vec{\kappa}_i|}$ and that R is the reflectance of each mirror, the intensity of the interference signal is:

$$I_{out} = E \cdot E^* = 2I_{in}R(1-R)\left(1 + cos\frac{4\pi\Delta d}{\lambda}\right) \tag{2.22}$$

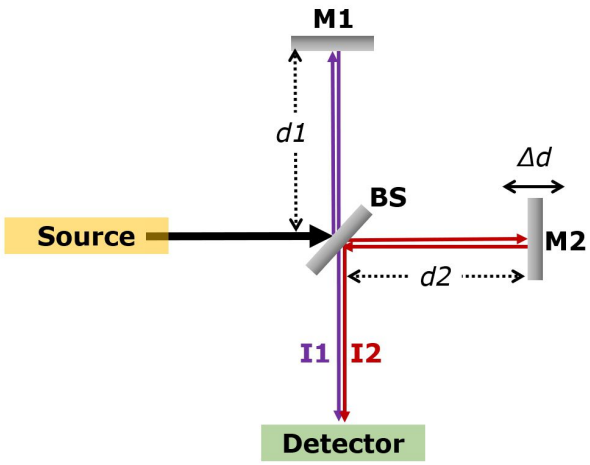

Figure 2.2: Schematic of the Michelson interferometer. The light source is split via a BS into two harms with intensity I1 and I2, respectively sent to mirrors M1 and M2 distant d1 and d2 from the BS. M1 is in a fixed position, while M2 is on a movable stage which allows scanning the interference of the two beams after they are recombined at the BS. Acronyms: BS - beam splitter; I1/I2 - intensity of the first/second harm; M1/M2 - mirror of the first/second harm; d1/d2 - length of the first/second harm; Δd - range of motion of the movable mirror M2.

Despite the Michelson interferometer has been conceived in the 1890s, only the advent of microcomputers able to perform Fourier transform made the spread of FT-IR spectroscopy possible in the 1960s [67]. The implementation of Michelson interferometer in commercial IR spectrometers gave rise to what we nowadays call *Fourier-transform infrared spectroscopy* or FT-IR. This technique has multiple advantages compared to dispersive IR spectroscopy. First of all, the possibility to acquire the full spectrum in a single scan provides the *multiplex advantage*, namely a faster acquisition rate compared to dispersive spectroscopy limited by the need to rotate the prism or grating to scan each wavelength. Moreover, FT-IR spectrometers record the whole spectrum at once, while the use of slits in dispersive spectrometers extensively reduces the intensity at the detectors, detrimental to the signal-to-noise ratio (SNR), called the *throughput advantage*. On top of that, FT-IR offers the *precision advantage*, namely the higher precision and resolution in wavelengths, with laser-based calibrations, compared to

dispersive devices based on external calibrations. This advantage allows comparing FT-IR spectra acquired at very different times. Overall, all the advantages combined allow to reach high SNR via averaging multiple scans maintaining a short acquisition time.

2.2.2 Fourier-transform infrared spectroscopy in transmission mode, FT-IR

The most common commercial FT-IR spectrometers work in transmission mode, meaning that the exciting beam passes through the sample under investigation and is partially absorbed [66, 68]. The attenuation of the beam from an initial intensity I_0 to I returns the sample transmittance:

$$T = \frac{I}{I_0} = 10^{-\varepsilon(\lambda)cl} \tag{2.23}$$

where the multiplication between the concentration c and the path-length l gives the number of molecules interacting with the exciting beam, while $\varepsilon(\lambda)$ is the molar extinction coefficient, a function of the optical properties of the sample as well as of the exciting wavelength. Equation 2.23 is shorten by calculating the log_{10} to:

$$A = log_{10}\frac{1}{T} = \varepsilon(\lambda)cl \tag{2.24}$$

which is the *Lambert-Beer law*, where A is the *absorbance* of the sample. This is additive, meaning that, for each wavelength, the total absorbance is the sum of the ones of each analyte in the sample. Of course, the validity of equation 2.24 is limited to a range of concentration that depends on the molecule: at too low concentrations the solvent will screen the molecules of interest (the analyte) leading to lower absorption signals than predicted by equation 2.24, while too high concentrations lead to similar screening effect by the analyte itself, again leading to a smaller absorption.

2.2.3 Attenuated total reflection, ATR-FTIR

Another approach for the acquisition of FT-IR spectra is in reflection [66, 68]. This exploits the reflection and refraction of a beam when it travels from a media with refractive index n_1 to one with index n_2 with an incident angle θ_i. If $n_2 < n_1$, for $sin\theta_i < n_2/n_1$ it is partially reflected with an angle $\theta_r = \theta_i$ and partially refracted into the medium with an angle θ_f according to the Snell's law:

$$\frac{sin\theta_i}{sin\theta_f} = \frac{n_2}{n_1} \tag{2.25}$$

However, at the critical angle θ_c, for which $sin\theta_c = n_2/n_1$, θ_f is 90° and the refracted beam will travel along at the interface of the two materials. For angles larger than the critical one, the refracted beam is evanescent and does not propagate. As a consequence, the radiation is only reflected into the first medium. This phenomena is called *total reflection* and is at the base of *attenuated total reflection Fourier-transform spectroscopy* (ATR-FTIR), also called *internal reflection spectroscopy*.

In total internal reflection, an evanescent wave with the same spectrum as the reflected one propagates within few micrometers beyond the interface with intensity diminishing

logarithmically. The interaction with the second medium allows recording its absorption spectrum in a measurement independent of the sample thickness. This technique allows to measure surfaces as well as optically thick samples and it is useful to reduce the contribution of highly absorbing solvents like water. Internal reflection elements (IRE) with a higher refractive index than the samples of interest (KRS-5, diamond, ZnSe, etc.) are used to achieve total internal reflection. The penetration depth into the sample can be estimated as:

$$d_p = \frac{\lambda}{2\pi n_{IRE}\sqrt{sin^2\theta_i - \left(\frac{n_s}{n_{IRE}}\right)^2}} \tag{2.26}$$

with n_{IRE} the refractive index of the IRE material and n_s the one of the sample. The fact that d_p is a function of the wavelength introduces a dependence of the signal intensity on λ not present in transmission mode. Moreover, the high variation of n_s around the central absorption frequency affects the penetration depth originating distorted bands compared to transmission spectrometers. This can be avoided by working at high incident angles, far from the critical angle. However, this is detrimental to the signal intensity. Therefore, the IRE usually has a trapezoidal shape which allows having multiple reflections (up to 25) to reach a high signal amplification.

2.2.4 Surface enhanced IR absorption, SEIRA

Surface enhanced infrared spectroscopy (SEIRA), which provides a higher sensitivity compared to common FT-IR spectroscopy, is usually applied on biomolecules, such as proteins [69]. This technique exploits the signal enhancement underneath the metal/analyte interaction, which has two origins. The first is the same as for similar techniques in the visible, such as surface-enhanced Raman scattering (SERS), and is based on the electromagnetic field enhancement due to surface plasma polariton (SPP) excitation of the nano-structures at the surface of the metal. However, it has been estimated that this enhances the signal only by a factor of 10, while the enhancement obtained in SEIRA is a factor of 100. Therefore, the second source of enhancement is far more important. This is called the *effective medium theory* and takes into account that the roughness of the surface is smaller than the IR wavelength which, therefore, probes a composite metal/sample medium, referred to as the "effective medium". The interaction between the IR radiation and the effective medium gives rise to an oscillating dipole of the absorbed sample coupled with the induced dipole of the metal. The dielectric of the metal is therefore altered enhancing the absorbance of the effective medium at the vibrational frequencies of the sample.

The different origin of the absorption signal induces major differences in the spectra acquired via SEIRA compared to FT-IR spectroscopy. The main difference is that, in the first, only the molecular vibrations orthogonal to the surface are strongly enhanced, while the ones parallel to it are extremely weak. SEIRA is strictly sensitive to the metal surface and represents a useful and increasingly spread technique for the analysis of structure and functions of biomolecules such as proteins at the level of a single monolayer.

2.2.5 Raman spectroscopy

Like infrared-based techniques, Raman is another spectroscopic method increasingly applied in similar studies to identify its potential for clinical applications [70, 71]. This technique explores the molecular vibrations of the sample analyzing the light scattered by the molecules, giving complementary information compared to FT-IR spectroscopy. In particular, it is based on two-photon processes involving excitation and detection of light at higher frequencies than FT-IR [64, 66].

The Raman interaction is due to the displacement of electrons and proton in the electric field of the incident light, which creates and induced dipole moment with intensity dependent on the applied electric field and on the deformability of the electron cloud, namely the molecular *polarizability* α:

$$\mu_{ind} = \alpha E \tag{2.27}$$

In the classical description of Raman spectroscopy, applying an electric field $E = E_0 cos(2\pi\nu t)$ will create an induced dipole emitting radiation in all directions with intensity given by $\alpha^2 E_0^2$. Therefore, the polarizability is the main factor in the Raman interaction with light. For small displacements, this can be expanded in a Taylor series:

$$\alpha = \alpha_0 + \frac{\partial \alpha}{\partial Q} Q + ... \tag{2.28}$$

where Q is the normal coordinate of the molecular vibration, given by $Q = Q_0 cos(2\pi\nu_v t)$, with ν_v the frequency of the associated normal mode. Higher orders are neglected in the harmonic approximation. Substituting equation 2.28 in equation 2.27, considering the definition of Q, the induced dipole can be written as follows:

$$\mu_{ind} = \alpha_0 E_0 cos(2\pi\nu t) + \frac{\partial \alpha}{\partial Q} \frac{Q_0 E_0}{2} [cos(2\pi(\nu - \nu_v)t) + cos(2\pi(\nu + \nu_v)t)] \tag{2.29}$$

This classical definition of the induced dipole moment highlights that three phenomena happen upon the interaction with light: the *Rayleigh scattering*, an elastic scattering dependent only on the frequency of the incident light, and two inelastic interactions leading to a different frequency than the incident light, called Stokes lines if they have lower energies ($\nu_S = \nu - \nu_v$) and anti-Stokes if they have higher energies ($\nu_{AS} = \nu + \nu_v$). Therefore, Raman spectra originate from inelastic interactions with light allowing the probing of the molecular vibrations ν_v. The anti-Stokes lines originate from higher vibrational states of the ground. The distribution of the population between vibrational states follows the Boltzmann distribution:

$$\frac{n_1}{n_2} = e^{(h\nu_v/kT)} \tag{2.30}$$

which explains why the anti-Stokes signals are usually less intense. Rising the temperature can increase their contribution relative to the Stokes lines. Equation 2.29 highlights one selection rule of Raman, which is that $\partial\alpha/\partial Q$ has to be non-zero. This means that only vibrations leading to a change in the molecular polarizability will be Raman active.

Similar conclusions can be derived from a quantum mechanic description analogous to the one followed for IR spectroscopy in section 2.1.2. It can indeed be shown that a similar matrix element found for IR can be obtained for Raman spectroscopy by replacing the molecular

electric dipole with the molecular polarization. For a mode with normal coordinate x, the matrix element has the following form:

$$\boldsymbol{\alpha}_{mn} = \alpha_0 \langle \chi_m | \chi_n \rangle + \left(\frac{\partial \boldsymbol{\alpha}}{\partial x} \right)_0 \langle \chi_m | x | \chi_n \rangle + \frac{1}{2} \left(\frac{\partial^2 \boldsymbol{\alpha}}{\partial x^2} \right)_0 \langle \chi_m | x^2 | \chi_n \rangle + \ldots \tag{2.31}$$

with χ_n and χ_m the initial and final vibrational wavefunctions. In off-resonance Raman spectroscopy, these states are in the electronic ground state, therefore $\langle \chi_m | \chi_n \rangle = 0$. As for IR spectroscopy and in agreement with the classical description, the selection rules for Raman spectroscopy are:

$$\left(\frac{\partial \alpha}{\partial x} \right)_0 \neq 0 \tag{2.32}$$

$$m - n = \pm 1 \tag{2.33}$$

The comparison of the first selection rule while the one in equation 2.19 for IR spectroscopy highlights that the IR-active modes are the ones associated with transitions that induce a change in the electric dipole of the molecule, while the Raman active modes are associated with a change in the molecular polarizability. For this reason, IR and Raman spectroscopy are considered complementary techniques. Resonant Raman spectroscopy, operated via tuning the exciting pulse to the electronic molecular excitation, offers the opportunity to obtain unique vibrational spectra by relaxing the second selection rule, in common between Raman and IR spectroscopy. This is possible because the initial and final vibrational states belong to two different electronic states. As for IR spectroscopy, surface-enhanced Raman spectroscopy (SERS) is also widely used for biological applications.

2.2.6 Field-resolved spectroscopy, FRS

In section 2.2.1, the main technology behind the most widely spread commercial FT-IR spectrometers has been introduced. These devices have several limitations, starting from the low brilliance of the thermal radiation sources to the limited dynamic range of the commercial detectors. Moreover, commercial FT-IR spectrometers are usually operated away from the shot-noise limit and are dominated by the detector noise [5]. Besides these technical limitations, one of the major drawbacks of common FT-IR spectroscopy is the strong water absorption of aqueous samples. To avoid this strong signal, in most of the studies on liquid biological samples (e.g. blood serum or plasma) these are first dried, leading to reproducibility problems due to the migration of macro-molecules to the periphery giving non-homogeneous distributions in the so-called coffee-ring effect [15, 58]. To overcome these limitations, a new laser-based technique has been developed in our laboratories, called *field resolved spectroscopy* or FRS. In this section, a description of the setup, with connected advantages and disadvantages, is discussed.

FRS set-up

One of the main limitations of FT-IR spectrometers is the low photon flux of the thermal sources. Better alternatives are the *quantum cascade lasers* (QCL), which give both pulsed and continuous laser radiation with about 10^4 times higher power density and are compatible with miniaturization [72]. Synchrotron sources provide a higher brilliance, but their availability and cost are too limiting. Other sources providing better performances are coherent mid-infrared (MIR) femtosecond broadband sources [38–41].

The source used in the FRS setup is the MIR radiation generated via difference frequency generation (DFG) driven by the compressed pulse of a Kerr-lens mode-locked ytterbium-doped yttrium-aluminum-garnet (Yb:YAG) thin-disk oscillator [38]. Figure 2.3 shows a schematic of the FSR setup [52]. The thin-disk oscillator provides a near-infrared (NIR) pulsed laser centered at 1030 nm with a repetition rate of 28 MHz, pulse duration (full-width half maximum) of 220 fs and an peak power of 14 MW. The NIR radiation is then broadened and compressed via self-modulation in bulk media (fused silica) using three consecutive Herriot-multi-pass cells able to provide a final pulse duration of 16 fs (Fourier limit of 15 fs). At this stage, the pulses cover from 920 to 1180 nm and have average power of 60 W. The MIR radiation is generated via intrapulse DFG focusing the compressed NIR pulses on an 11 mm thick type I $LiGaS_2$ (LGS) crystal. The generated MIR radiation, with a spectral coverage that spans from 980 to 1550 cm^{-1}, passes through a chopper (7.5 kHz) and is focused with a spot diameter of 420 μm and average power of 50 mW on the liquid cuvette in ZnSe (2 mm thickness, *Micro Biolytics GmbH*). The cuvette is filled via an automated microfluidic system suited for the measurement of liquid biological samples, programmed to exchange aqueous samples and water reference. Active noise eater based on an acousto-optic modulator device, as well as the lock-in detection via chopping the MIR signal, are employed to further improve the stability of the system.

Employing sources with higher MIR power and brilliance does not automatically lead to better sensitivity because of limitations due to intensity noise and dynamic range [52]. To overcome the lasts, the detectors commonly used in commercial FT-IR spectrometers are replaced with the field resolved *electro-optic sampling* (EOS), which allows the measurement of the electric field of the vibrating molecules. The high waveform stability of the MIR pulses together with the high dynamic range and the field-scaling of the signal allows the precise subtraction of the signal from the water reference thus isolating the signal of the other molecules [52, 73], overcoming the limitations of drying the samples. The time-resolved trace obtained after the subtraction is called *electric-field-resolved molecular fingerprint* (EMF).

This detection scheme uses a short sampling pulse to probe the electric field of interest. In particular, the NIR is separated by the MIR after the DFG crystal via a dichroic mirror, attenuated and used as sample pulse in the EOS detection. The two beams are spatially recombined in a germanium plate, at Brewster angle for the MIR beam, and combined in a 96 μm-thick GaSe crystal via sum frequency generation (SFG). This allows recording the interference with the sampling pulse via balanced heterodyne based on two photodiodes. A short-pass filter is used to enhance the SNR. The balancing is obtained via a half-wave plate and a quarter-wave plate. The detection is confined in the temporal window of the propagation of the NIR in the EOS crystal. This temporal gating in the detection allows scanning the MIR signal by tuning the delay with the sampling pulse, tracked via an interferometric delay tracking system (IDT).

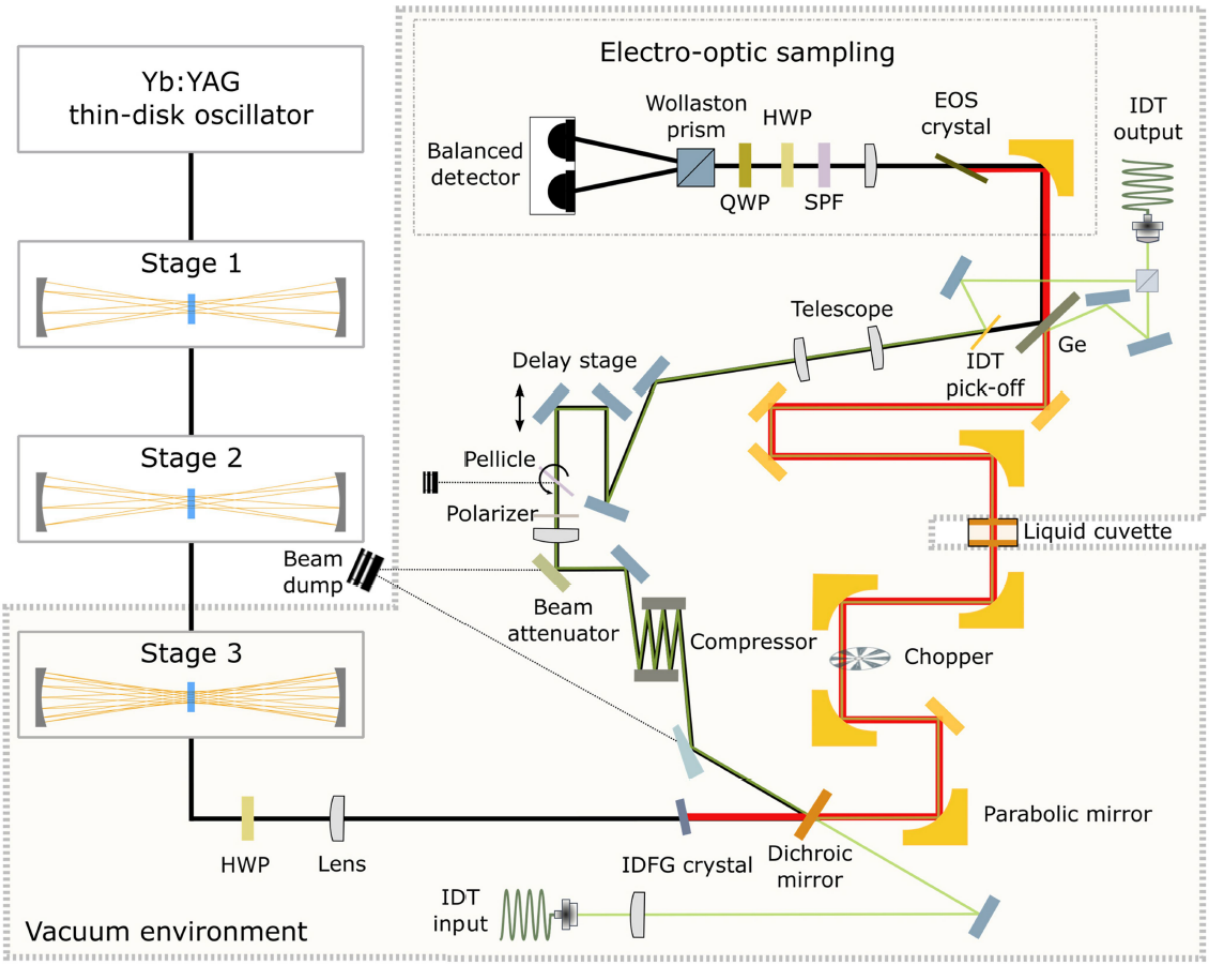

Figure 2.3: Scheme of the FRS setup used in this dissertation. The source is a Yb:YAG thin-disk oscillator. The 1030 nm radiation (full-width half maximum of 220 fs) is then broadened and compressed via three consecutive Herriot-multi-pass cells (Stage 1, Stage 2 and Stage 3) providing a final pulse duration of 16 fs (Fourier limit of 15 fs) with spectral coverage spanning from 920 to 1180 nm and have an average power of 60 W. The MIR radiation is generated via intrapulse DFG (IDFG) in a 11 mm thick type I LGS crystal. The generated MIR radiation has a spectral coverage that spans from 980 to 1550 cm^{-1}. After passing through a chopper, the MIR is focused on the liquid cuvette in ZnSe (2 mm thickness, *Micro Biolytics GmbH*) with an average power of 50 mW. The NIR generated at the LGS crystal is separated by the MIR and used as the sampling pulse in the EOS detection. The MIR and NIR beams are spatially recombined in a germanium plate and combined in a 96 μm-thick GaSe crystal via SFG. A short-pass filter (SPF) is used to enhance the SNR. The balancing is obtained via a half-wave plate (HWP) and a quarter-wave plate (QWP). The delay between the two pulses is tracked via an interferometric delay tracking system (IDT). Acronyms: Yb:YAG - ytterbium-doped with yttrium aluminium garnet ($Y_3Al_5O_{12}$); MIR - mid-infrared; NIR - near-infrared; IDFG - intra-pulse difference frequency generation; LGS - lithium gallium sulfide ($LiGaS_2$); EOS - electro-optic sampling; SFG - sum frequency generation; SPF - short-pass filter; SNR - signal-to-noise ratio; HWP - half-wave plate; QWP - quarter-wave plate; IDT - interferometric delay tracking system. Adapted by permission from Copyright Clearance Center: Springer Nature Limited, Nature [52] (Field-resolved infrared spectroscopy of biological systems, Ioachim Pupeza et al., 2020).

FRS: advantages and further developments

One of the main advantages of FRS spectroscopy is the waveform-stability of the MIR pulses which allows subtracting the signal of a reference measured a few minutes before from the samples of interest, thus suppressing the strong absorption of water allowing the measurement of biosamples in their natural aqueous environment. The high brilliance source boosts the performance of commercial FT-IR spectrometers. However, this is advantageous only if combined with a detection system able to cover a high dynamic range. In this respect, the temporal gated detection scheme is advantageous as it allows the detection of intense signals that would saturate the detector if the whole signal would reach it at once. The EOS detection is relatively similar to the interferometric approach employed in FT-IR spectroscopy. In both cases, the arm encoding the signal of the sample is detected using a second arm acting as the local oscillator for the heterodyne/homodyne detection. However, in FT-IR spectroscopy the molecular signal is strongly affected by the technical noise of the exciting source, as it builds upon it. On the other hand, the impulsive excitation combined with the temporal gated detection in FRS allows to temporally filter the exciting pulse thus isolating an "excitation/background-free" signal. The impact of the MIR noise is reduced making only the noise of the sampling pulse relevant, which is operated close to the shot-noise limit.

Overall, FRS offers "background-free" measurements of the coherent response of impulsively excited molecules circumventing the limitations imposed by source-noise and detector saturation. The combination of a high brilliance MIR source and the sensitive EOS detection allows reaching the limit of detection (LOD) of 200 ng/mL with the described FRS set-up, about 40 times lower than FT-IR (8 $\mu g/mL$) and 5 orders of magnitude lower than the most abundant molecule in human serum (albumin). In other words, this implies that FRS can potentially detect more low abundant molecules compared to the commercially available FT-IR spectrometers, without the need for concentrating, fractionating or depleting the biosamples. The analysis of intact systems with high optical and physical thickness has been demonstrated in [52] and highlights the potential of FRS spectroscopy for biological, biomedical, pharmaceutical and ecological applications. In this dissertation, FT-IR and FRS measurements are performed on full samples preserving the natural aqueous environment of blood-based biofluids.

Further developments are currently being implemented, such as the development of sources with super-octave spectral coverage and higher intensity dynamic range, potentially able to reach LOD below 50 ng/mL and faster scanning operations able to freeze the excitation pulse fluctuations in each measurement. Higher day-to-day reproducibility is also necessary, due to the operation of the whole system from the third Herriot cell below 1 $mbar$ to reduce the water absorption. This indeed creates a strong tension on both sides of the liquid cuvette with unwanted effects on the measurement reproducibility, as tackled in the next chapter.

2.3 Data analysis

The tools used in this dissertation for the analysis of FT-IR and FRS data are briefly introduced in the following section. In particular, all dimensionality reduction and machine learning analysis have been carried using python version 3.7.3 [74, 75], while the matching of cases and controls cohorts has been performed using RStudio version 1.3.1093 [76].

2.3.1 Principal component analysis

The measurement of large cohorts of individuals via FT-IR and FRS result in big datasets, however, most of the spectral information is not informative. Therefore, dimensionality reduction techniques able to retain the most informative part of the data are essential for any further analysis. There are different ways to perform it, such as *feature selection*, which simply retains a subset of the original features, and *subspace projection*, which constructs new representations with lower dimension as linear combinations of the features of the original data. In this dissertation, dimensionality reduction is performed via the key instrument of the subspace projection approach, namely *principle component analysis* (PCA). In the following chapters, PCA has been applied to the data before any other analysis, except for deriving the SVM coefficients. A more detailed description of PCA can be found in [77].

Dimensionality reduction has multiple advantages. To start with, it reduces the computational time and power consumption, boosting the performances by retaining only the features encoding useful information. It also reduces the tendency to overfit the data of supervised learning algorithms. Moreover, dimensionality reduction is useful for feature extraction and/or visualization purposes, for which PCA is commonly adopted.

The algorithm for the PCA unsupervised learning model is hereby reported. This is based on the assumption that the dataset can be modeled as multivariate stochastic observations with Gaussian distributions, according to which the covariance matrix of the data suffices to determine the optimal projection subspace. Let $\{x_1, x_2, ...x_N\}$ be the N-dimensional training dataset, each x_t being an M-dimensional vector, with covariance matrix $\mathbf{R} = \{\hat{r_{ij}}\}$ and:

$$\hat{r_{ij}} = \frac{1}{N}\Sigma_{t=1}^{N} x_t^i x_t^j \tag{2.34}$$

with $i, j = 1, 2, ...M$. Applying the spectral decomposition on $\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ the eigenvectors $\mathbf{V}$ composing a new basis set provide the *principal components*:

$$\mathbf{y}_t = \mathbf{V}_m^T \mathbf{x}_t \tag{2.35}$$

with $\mathbf{V}_m$ the first m eigenvectors of $\mathbf{V}$. Therefore, only the first m "principal eigenvectors" are considered to reduce the dimensionality from M to m. The elements of the diagonal matrix $\mathbf{\Lambda}$ are the corresponding eigenvalues which monotonically decrease from λ_1 to λ_M.

The algorithm simply describes the projection of the data along the principal components exploiting the statistical dependence and inherent redundancy embedded in the training data to derive a more compact but highly representative dataset. Under the Gaussian distribution assumption, it can be demonstrated that the PCA is well representative of the original dataset because it satisfies two criteria. The first is the *mean-square-error criterion* which shows that the PCs are the best estimates of the original data, namely that the error $\epsilon(\mathbf{x}|\mathbf{z}) = \min_{\mathbf{y}\epsilon\mathbb{R}^m} ||\mathbf{x} - \hat{\mathbf{x}}_y||$,

with $\hat{\mathbf{x}}_y$ the best estimate of x from the principal component $\mathbf{y}$. This is satisfied by the following theorem about the optimality of PCA in reconstruction error: PCA offers an optimal subspace projection with the minimal expected value of reconstruction error:

$$\epsilon[||x - \hat{x}_y||^2] = tr\{\mathbf{R}_X\} - \Sigma_{i=1}^{m}\lambda_i = \Sigma_{i=m+1}^{M}\lambda_i \tag{2.36}$$

where the trace $tr\{\mathbf{R}_X\}$ denotes the sum of all the diagonal elements of the matrix. This states that the PCA retain the while information of the original data retaining the most valuable in the first m components.

Another criterion is the *maximum-entropy criterion*, based on how much information of the original data is retained in the reduced-dimension vector $\mathbf{y}$. The entropy is a measure of the amount of information in a random vector. It can be shown that maximizing the mutual information $I(\mathbf{x}|\mathbf{y})$ between the dataset $\mathbf{x}$ and the principal components $\mathbf{y}$ is equivalent to maximizing the entropy of the PCA, $H(\mathbf{y})$. Being $\mathbf{y}$ a subspace of $\mathbf{x}$, it cannot contain more information, therefore $I(\mathbf{x}|\mathbf{y}) = H(\ \mathbf{y})$ and $H(\mathbf{y}) \leq H(\mathbf{x}|\ \mathbf{y}) = H(\mathbf{x})$. Ideally, $H(\mathbf{y})$ should be as close as possible to $H(\mathbf{x})$. Being the probability distribution of $\mathbf{x}$ and, therefore, of $\mathbf{y}$ Gaussian:

$$p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^m|\mathbf{R}_y|}} exp\Big\{ -\frac{1}{2}(\mathbf{y}^T\mathbf{R}_y^{-1}\mathbf{y})\Big\} \tag{2.37}$$

and from the definition of entropy:

$$H(\mathbf{x}) = -\int p(\mathbf{x}) log[p(\mathbf{x})] d\mathbf{x} \tag{2.38}$$

it can be shown that PCA offers an optimal subspace projection with maximal mutual information between $\mathbf{x}$ and $\mathbf{y}$:

$$I(\mathbf{x}|\mathbf{y}) = \frac{1}{2}\Sigma_{i=1}^{m} log_2(2\pi e\lambda_i) \tag{2.39}$$

There are different numerical methods to compute the PCs, such as *singular value decomposition* (SVD) and *spectral decomposition*. In this dissertation, SVD is selected. The SVD algorithm is applied directly on the matrix data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n]$ in the following way:

$$\mathbf{X} = \begin{cases} \mathbf{V}\begin{bmatrix}\mathbf{D} & \mathbf{0}\end{bmatrix}\mathbf{U} & \text{if N} \geq \text{M} \\ \mathbf{V}\begin{bmatrix}\mathbf{D} \\ \mathbf{0}\end{bmatrix}\mathbf{U} & \text{if M} > \text{N} \end{cases} \tag{2.40}$$

with $\mathbf{V}$ and $\mathbf{U}$ respectively $M \times M$ and $N \times N$ unitary matrices and $\mathbf{D}$ a diagonal matrix of dimension $M \times M$ or $N \times N$ based on the $min\{M, N\}$. The PCA representation will be again defined as $\mathbf{y} = \mathbf{V}_m^T\mathbf{x}$ with $\mathbf{V}_m$ the $m \times N$ matrix from the first rows of the matrix $\mathbf{V}$.

In the general expression of principle components (equation 2.35), the eigenvectors of the matrix $\mathbf{V}$ are just projections, without any information regarding the amount of variance explained by each PC which is expressed by the eigenvalues in $\mathbf{\Lambda}$. The multiplication of the eigenvectors by the square root of the corresponding eigenvalue returns the *loading vector*. While PCA separates the covariance matrix into explained variance and direction, the loading vectors construct back that part of information of the original data, namely the covariances between the original variables and the components. These will be useful in the next chapters to identify the frequency components responsible for the largest spectral variability.

2.3.2 Machine learning algorithms

In this dissertation supervised and unsupervised algorithms are used in the same fashion for both the FT-IR and FRS data. Here a quick overview of the methods applied for the reported analysis is presented. A thorough discussion about the reported methods is available in dedicated books [78, 79].

Supervised binary classifications: support vector machine, SVM

Supervised machine learning algorithms "learn" the data to build a model for the distribution of classes on a *training set*, based on the known labels, and perform predictions on unlabeled data constituting the *test set*. One of the multiple applications is data mining, boosting the robust classification of two or more classes in a given dataset. This section focuses on one of the most spread supervised algorithms, *support vector machine* or SVM, which is applied after PCA on both FT-IR and FRS data for the binary classification of individuals based on their phenotypes.

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N\}$ be a dataset and $Y = \{y_1, y_2, ...y_N\}$ the corresponding categorical labels or *teacher values*. The input dataset would be $[X, Y] = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_N, y_N)\}$. In linear regression analysis, which models the dependence of a dependent variable $\mathbf{y}$ from an independent variable $\mathbf{x}$, the aim is to find a vector $\mathbf{w}$ and the intercept (or bias) b that can approximate the dependent variable, namely $\mathbf{w}^T\mathbf{x}_i + b \approx y_i$. While in linear regression this condition is applied to all datapoints, SVM takes into account only the most meaningful of them, called *support vector*. In particular, for SVM the equality becomes the inequality:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b - y_i) \leq 0 \quad \forall i = 1, ...N \tag{2.41}$$

or, in other words, it has to satisfy $\mathbf{w}^T\mathbf{x}_i + b \geq +1$ and $\mathbf{w}^T\mathbf{x}_i + b \leq -1$. In particular, in an ideal case where the two classes are completely and linearly separable, $\mathbf{w}^T\mathbf{x}_i + b = \pm 1$ are the two *marginal hyperplanes* or *supporting hyperplanes*, equidistant from the *decision hyperplane* given by $\mathbf{w}^T\mathbf{x}_i + b = 0$. The data on the marginal hyperplanes are the *support vectors*, while the rest are called *non-support vectors*. This approach, called the *maximal margin classifier*, aims at defining the marginal hyperplanes with the largest mutual distance. However, this works only in the ideal case of a good separation, which is most of the time not the case. Simply allowing a few miss-classifications gives *soft margins* which work better for the classification of the whole dataset in not-ideal settings. In this case, the support vectors are the observations on the edges and within the soft margins.

Soft margin classification is an example of the bias/variance trade-off: it increases the bias to lower the variance. In particular, the $bias^2$ measures the error on the training data, while the error on the test is $bias^2 + variance$ and will, therefore, always be higher. However, for a given model, increasing the number n of datapoints the $bias^2 + variance$ increases and the $bias^2$ decreases, converging for $n \to \infty$. Therefore, the number of datapoints, in our case of spectra or time traces, influences the performance of the classifier (see section 4.2.3). Moreover, the $bias^2 + variance$ and $bias^2$ depend on the complexity of the model. Models with low complexity lead to a *underfitting* of the training data, while increasing the complexity they can fit better the training set, therefore decreasing the $bias$. However, if the level of complexity is too high, they will perform poorly on the test (high $variance$) leading to strong *overfitting*. The bias/variance trade-off means finding the optimal complexity model able to minimize both errors.

In k-fold *cross-validation* (CV), the training set is split into k (usually 10) subsets: k-1 subsets are used to train the model which is tested (validated) on the remaining subset. This procedure is repeated using different subsets as a test and the performance is the average of the values computed in the loop. Cross-validation can be computationally expensive, but each training does not exclude too much data, which is advantageous for small datasets. The classification will address correctly some of the positive (true positive, TP) and some of the negative (true negative, TN) datapoints and it will wrongly identify other positive (false negative, FN) and negative (false positive, FP) ones. A way to represent the classification performance and derive a unique number is using the *receiver operating characteristic* (ROC) curves, obtained by plotting the true positive rate (TPR = TP/(TP + FN)) against the false positive rate (FPR = FP/(FP + TN)) at different classification thresholds. The efficiency of a binary classification can be determined by the *area under the curve* (AUC) of the ROC curve, which goes from 50% for a random classification (no separation between the two classes) to 100% for ideal separations. Besides giving a compact estimate of the classification efficiency, AUC is also advantageous because it is scale-invariant, measuring how well predictions are ranked independently from their absolute values, and is classification-threshold-invariant.

Linear SVM has been found to perform better than non-linear classification algorithms on both FT-IR and FRS data of human blood-based biofluids and is therefore applied after PCA in 10-fold cross-validation (repeated 10 times) for all binary classifications in this dissertation. Moreover, for linear SVM, the magnitude of the feature weights, called *SVM coefficients*, indicates the relevance of each feature for the corresponding binary classification [80]. These coefficients are calculated directly on the spectral and temporal data. The SVM binary classifications on the EMFs in the time domain have been performed by scanning increasingly narrower time windows, selecting the thickness giving the highest classification efficiency and scanning it again to identify the temporal range giving the highest AUC. This *optimal time window* has been used to calculate the SVM coefficient in the time domain. The distance between the average AUC obtained for the training set from the standard deviation of the test set has been used to evaluate the potential overfitting. SVM is usually applied after *standard scalers* which standardize the features by subtracting the mean μ and scaling to unit variance s ($z = (x - \mu)/s$). However, when applied on FT-IR and FRS data standard scalers lead to overfitting, especially for small cohorts, and are therefore not applied in this dissertation.

Unsupervised algorithms

Unsupervised algorithms are widely used to identify *cluster* of data with a similar structure without apriori knowledge, namely without taking into account any given label and are therefore called *unsupervised clustering algorithms*. Among the multiple applications of clustering, they can be used to gain a better understanding of the structure of complex datasets by summarizing the distribution of specific parameters among the identified clusters. Their performance depends on several factors, such as the number of clusters, the topology of node vectors, the objective function for clustering, the type of clustering algorithms, the initial conditions and the evaluation criterion for selecting the best clustering.

Before going into the details of the unsupervised algorithms used in this dissertation, it is useful to introduce a method for the definition of the optimal number of clusters. One way to proceed, used in the following chapters, is by adopting the *elbow method* which plots the

within-cluster sum of squares (WCSS) clustering the data for an increasing number of clusters. Of course, WCSS will be maximum for one cluster (the whole data set) and will decrease to zero for the maximum number of clusters possible, namely if the number of clusters equal to the number of datapoints, as in this case each cluster will count only one point. The optimal number of clusters is the one for which the slope of the WCSS changes abruptly, called "the elbow".

The clustering algorithms can be based on different approaches: *centroid clustering*, based on the identification of the center of the clusters, *density clustering*, based on the density distribution of the datapoints, *distribution clustering*, which assume that the data follow pre-selected distributions, and *connectivity clustering*, based on the mutual distance of the datapoints. One of the most commonly used centroid clustering algorithm is *K-means*, which, selected the number K of clusters, chooses k random centers, allocates each data point to the cluster whose center is nearest, ensuring that every cluster has at least one datapoint. It further replaces the cluster centers with the mean of the elements in their clusters iteratively until it reaches the smallest sum of squared distances of each point x_i to the center c_j of the respective *j-th* cluster:

$$\Phi(\delta, \mathbf{c}) = \Sigma_{i,j}\delta_{i,j}[(x_i - \mathbf{c}_j)^T(x_i - \mathbf{c}_j)] \tag{2.42}$$

where $\delta_{i,j}$ is 1 if x_i belongs to the *j-th* cluster and 0 otherwise, therefore acting as a switch. If we knew where the center of each of the clusters was, it would be easy to define $\delta_{i,j}$, and the other way around. Iteration helps finding the $min(\Phi(\delta, \mathbf{c}))$.

An alternative method among the connectivity clustering algorithms is called *agglomerative clustering*. In this case, instead of the distance between datapoints and centroids, the mutual distance between datapoints is considered and plotted in a dendrogram. In this case, each point is first assumed to be an independent cluster and is associated with the closest datapoint, therefore creating a hierarchical clustering. The height at which to truncate the dendrogram will determine the number of clusters and strongly depends on the data. Different approaches, such as the elbow method, can be helpful to determine the number of clusters.

The clustering methods discussed so far are based on mutual distances in the feature space. The algorithms based on distribution clustering, instead, consider a dataset as a collection of different clusters, each based on probability models. Knowing the model would make it possible to attribute each datapoint to its cluster, as well as, knowing to which cluster each point belongs, would allow to easily define the probability model. This is similar to the issue seen for K-means, for which one should know both the centers and to which center each datapoint belongs. As for K-means, also in these clustering algorithms, the issue is solved by iterative calculations in what is called *expectation maximization* or EM. The distribution clustering algorithms start by defining random components, such as random centers obtained via k-means, and then calculate the probability for each data point of being generated by the corresponding probability model. It then iterates the optimization of the parameters to maximize the likelihood of each assignment until it converges to a local optimum. An important example of distribution clustering algorithms is the *Gaussian mixture model* (GMM), based on the assumption that the datapoints are generated from a mixture of Gaussian distributions. GMM uses the EM algorithm to fit mixture-of-Gaussian models.

A way to compare the clustering performances of different distribution clustering algorithms for different parameters (e.g.: number of clusters, probability distribution, etc.) is via adding penalties for increasing model complexity to the training error. The $bias^2$ alone, indeed, is not a

good measure since it will decrease at increasing model complexity leading to a higher *variance*. Adding a penalty will identify an "optimal model complexity" which satisfies the bias/variance trade-off. There are two comparable methods to define the penalty: the *Akaike information criterion* (AIC) and the *Bayesian information criterion* (BIC). In both cases, the penalty is the *log-likelihood L*, which increases for a better model. Both methods aim at minimizing the penalized log-likelihood, in particular $2k - 2L$ in AIC and $2k(logN) - 2L$ in BIC, where k is the number of parameters.

To compare the efficiency of different clustering algorithms, the silhouette score is a good solution. This is calculated for each sample based on the mean intra-cluster distance (I_c) and the mean nearest-cluster (N_c) distance, namely the distance between the sample and the nearest cluster it does not belong to, as $(N_c - I_c)/max(I_c, N_c)$. The silhouette score can go from 1 for a perfect assignment to -1 if the sample belongs to another cluster. If the score is 0, the two clusters overlap and it is difficult to attribute the sample to any of the two clusters.

In this dissertation, unsupervised methods are applied on FT-IR and FRS data to identify clusters on all PCs encoding 99.9% of the total variance. The elbow method is used to identify the optimal number of clusters, namely the one leading to a drastic change in the slope of the WCSS and the best performing algorithm is selected via their silhouette scores. The clustering methods applied are: agglomerative clustering (with affinity = 'euclidean' and linkage = 'ward', as well as with affinity = 'cosine' and linkage = 'average'); K-means; Gaussian Mixture Model (with covariance type = 'full', as well as with covariance type = 'spherical'). The BIC scores are also used to address the best performing GMM clustering for a different number of components and covariance types (full and spherical).

2.3.3 Case-control matching algorithm

In this dissertation, the IR spectroscopy of biosamples is performed to identify the vibrational signatures of common phenotypes via binary classifications with control cohorts. However, to isolate phenotype-specific signatures, other parameters (e.g. age, gender, comorbidities, etc.) need to be equally distributed between cases and controls to guarantee that the target phenotype is the only difference between the two groups. One way to do so is by random sampling, which increases the chances that other parameters are balanced. However, if a parameter correlates with the target phenotype, it will not be equally distributed between cases and controls. Different correlations, as is the case for different studies (e.g. cross-sectional and case-control cohorts), will therefore lead to different fingerprints for the same phenotype. To overcome this issue and enhance the phenotype-specificity, one solution is to match cases and controls for the known parameters [81].

One way to match two cohorts is via *propensity score* [82], which are the likelihood of being positive ($i = 1$) to the target phenotype h based on a set of covariates X ($h = i|X$), which can be derived via logistic regression. Therefore, controls with the same likelihood of being positive to the phenotype can be considered as a match for the cases with a similar propensity score, even though their covariates might differ. In other words, after matching via propensity score, the covariates will be balanced on average. Summarizing many covariates in a single score is a major advantage as compared to matching the known covariates one by one since the last would return fewer cases the more parameters are considered in the matching.

Matching can then be performed using *optimal full matching*, which minimizes the total

distance between cases and controls with comparable propensity scores matching one or more individuals to each case, making it ideal to avoid discarding cases. The matching applied generates a distance matrix based on the propensity scores. One way of calculating it is via the Mahalanobis metric distance, namely as $d(i, j) = (u - v)^T C^{-1}(u - v)$, whit u and v the matrices with the covariates of cases and controls respectively and C the variance-covariance matrix of the controls. To make sure that the matched control is not too far in the multidimensional space, which is more likely for an increasing number of covariates, the distance is estimated based on a specified tolerance, or *caliper* ε, such that $||d(i, j) < \varepsilon||$.

In this dissertation, full matching based on propensity score and Mahalanobis distance has been used to match cases and controls using RStudio (version 1.3.1093) [76].

Chapter 3

Technical and biological noise of FT-IR and FRS fingerprints

The development of any disease alters the natural physiological state of specific organs and tissues and, therefore, of the biofluids in contact with them. This makes human blood serum and plasma, which access all organs in our body, particularly unique and powerful for applications in health monitoring. Their chemical composition can be investigated in a fast and quantitative fashion via infrared (IR) spectroscopy, which gives a snapshot of their molecular concentrations and structures. However, multiple factors could potentially influence their composition and, therefore, their IR signature, including demographic and other common parameters. As a consequence, the standard deviation of spectra between different individuals, later on referred to as the *between-person spectral variability*, is very large [58, 60, 83]. This has been often addressed as one of the major confounding factors challenging the reliability of IR spectroscopy for disease diagnosis [61]. However, so far there has been no major effort in the robust characterization of the signatures and origin of such strong variability in any large study.

Samples from about 2100 KORA-FF4 individuals have been measured via FT-IR and FRS spectroscopy. From the Lasers4Life (L4L) clinical study, only the FT-IR spectra of the 620 cancer-free individuals have been selected for comparison. In this section, principal component analysis (PCA) and support vector machine (SVM) binary classifications, as well as unsupervised machine learning algorithms, are used to address the impact of age, gender, body mass index (BMI), smoking status, alcohol consumption and inflammation on the between-person spectral variability of FT-IR and FRS data recorded for the healthiest individuals of KORA-FF4 and L4L cohorts, namely the individuals non-symptomatic (NSP) for the know common medical conditions and with normal glucose tolerance (NGT). Before addressing the biological variability among NSP/NGT individuals, the noise intrinsic to the measurement of liquid biological samples is investigated in the first part of this chapter using the KORA-FF4 measurement campaign, the largest ever performed via FRS spectroscopy.

3.1 Technical noise characterization

The KORA-FF4 and L4L cohorts are independent, implying that different clinical equipment and standard operating procedures (SOPs) for blood drawing, sample handling and quality check have been adopted. The lack of standard SOPs has been often addressed as the major source of preanalytical errors in clinical practice [58, 61, 84, 85]. However, the inter-clinical impact on FT-IR spectra of human blood bio-fluids has been shown negligible compared to the between-person spectral variability, which includes the biological information we seek in our analysis [58, 60, 83]. This not only allows the comparison of independent cohorts but, more importantly, it opens the way for IR spectroscopy applications for health monitoring. The clinical and demographic parameters collected for KORA-FF4 and L4L differ in content and source. In both studies, each phenotype is self-reported. The only exceptions are in KORA-FF4 for diabetes, identified via oral glucose tolerance test (OGTT), and heart attack, involving individuals who had suffered from a myocardial infarction. Because of the differences between the two cohorts, any comparable outcome is expected to be a signature specific to the phenotype under investigation.

The samples have been shipped to our laboratories on dry ice. Once arrived, several aliquotes have been prepared and stored at -80°C until measurement. The aliquotes of L4L have been prepared manually, while the numerous samples of KORA-FF4 have been processed via an automated liquid handling system. The SOP applied for the sample preparation was the same for FT-IR and FRS measurements: the samples are randomized, thawed in a water bath, mixed with a vortex mixer for about 30 s and spun down for about 1 min. Replica of the same human blood serum sample have been used as quality control (QC, BioWest, Nuaillé, France) and measured every 5 samples to check for instrument drifts and confirm that chemical changes of the biosamples are negligible in the 3-4 h time span of the measurements [86]. About 450 QCs have been measured during the KORA-FF4 measurement campaign. The QCs have a similar chemical complexity compared to the actual samples. Therefore, the standard deviations of the FRS time traces and absorption spectra of QCs provide a useful characterization of the technical noise of the respective technique when the last is applied on human blood serum or plasma and is suitable for preprocessing optimization, as discussed in the following.

3.1.1 Preprocessing optimization: FT-IR

The FT-IR spectra of human blood serum and plasma have been performed in their natural aqueous environment via MIRA-Analyzer (micro-biolytics GmbH), a state-of-the-art spectrometer dedicated to the measurement of liquid biological samples. The transmission cell in CaF_2 is 9.6 μm thick. The spectral coverage spans from 1000 to 3000 cm^{-1},truncated at 1500 cm^{-1} when compared to FRS. The implemented software alternates sample and water measurements at room temperature with a resolution of 4 cm^{-1}. The IR spectrum of water is subtracted from the signal of the corresponding sample introducing negative absorption values, which are set back to zero by adding a standard water spectrum to flatten the silent region (1850 - 2300 cm^{-1}) [87, 88]. The normalization of each spectrum to its mean area reduces the technical noise by a factor of 26 and, consequently, the standard deviation of the spectra of the samples by a factor of 8 (Figure 3.1). Therefore, unless specified, all FT-IR data in this work are normalized.

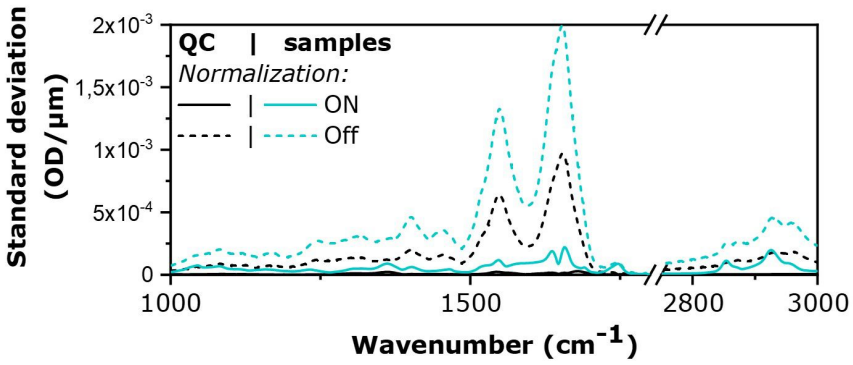

Figure 3.1: FT-IR spectral standard deviation of the 450 QCs and the 2100 samples measured during the KORA-FF4 measurement campaign. The comparison between not normalized data (dotted lines) and the data normalized to their mean area (solid lines) highlights that normalization reduces the technical noise by a factor of 26 on QCs and of 8 on the samples. Acronyms: QCs - quality controls.

3.1.2 Preprocessing optimization: FRS

As for FT-IR, the FRS measurements of human blood serum and plasma are performed on liquid samples. In particular, 12 time-traces are recorded from -0.4 to 7 *ps* and averaged for every measurement. After each sample, a water reference is recorded and subtracted directly in the time domain to remove the strong water absorption. The subtraction of the average time-trace of the water reference from the one of the respective sample isolates the coherently emitted *electric-field-resolved molecular fingerprint* (EMF) of the excited biomolecules. The strong noisy residual of the exciting pulse around the zero of the EMFs can be filtered out using a *high temporal pass filter* (HTPF) to isolate the background-free molecular fingerprint, as described later in this section. The raw time traces are subjected to technical noise that covers the biological information we seek and need to be compensated. To this end, as done for FT-IR in the previous section, the EMFs of all 450 QCs measured with the KORA-FF4 samples are used to evaluate the technical noise intrinsic to the measurement of biological samples. In this section, the effect of each preprocessing step is evaluated to identify the combination able to minimize the technical noise.

Despite the high stability of the laser, reproducing every day the same experimental settings it is highly challenging. One if the consequence it that the centers of the raw time-traces are slightly different from day to day (Figure 3.2b). The *principal component analysis* (PCA) reveals how this technical noise introduces a dependence on the measurement day (Figure 3.2a). Tho compensate for it, several centering options have been investigated. In the following, *Hilbert centering* is applied. This uses the Hilbert transformation to retrieve the envelope of each reference pulse and centers their maximum to a common zero. The same shift along the time axis is applied to the average time-trace of the corresponding samples before the subtraction (Figure 3.2c, d).

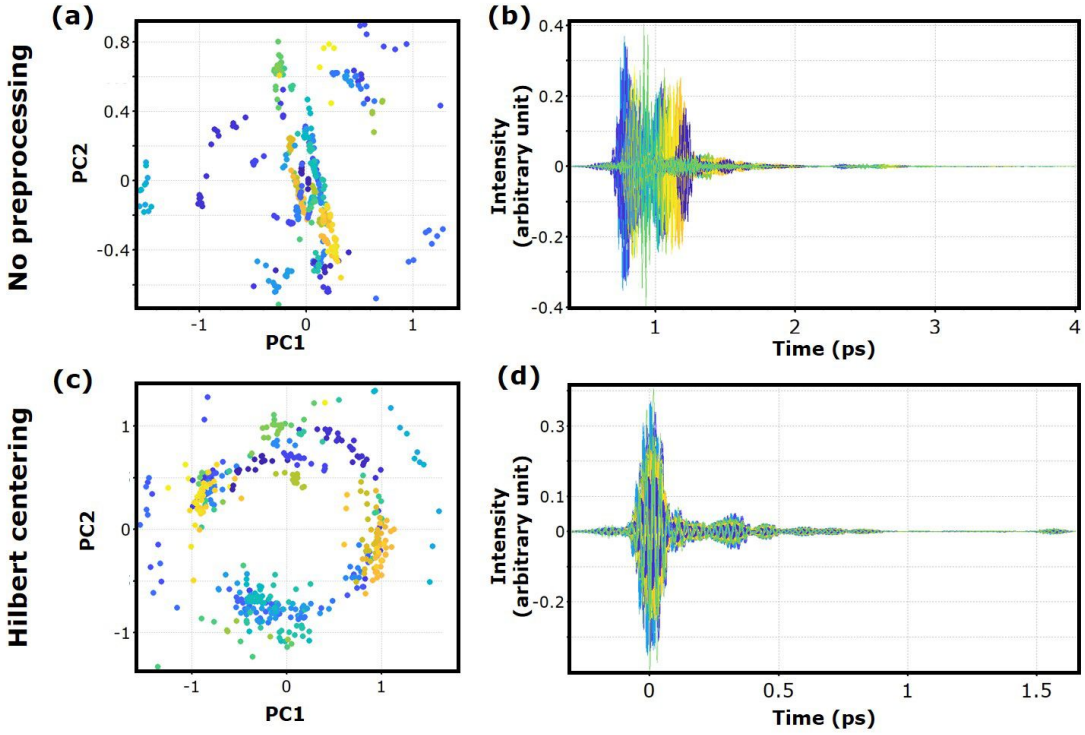

Figure 3.2: FRS preprocessing optimization in the time domain: EMFs centering. The EMFs recorded for the 450 QCs are used to test several preprocessing steps and to evaluate the technical noise of FRS measurements on blood-based biofluids. The color gradient highlights the measurement day. (a) The PCA analysis of (b) the raw EMFs highlights a day-to-day dependence due to different centering of the time traces. After applying the Hilbert centering, (c) the PCA shows a different day-to-day dependence and (d) the EMFs are more similar to each other. Acronyms: FRS - field-resolved spectroscopy; EMF - electric-field-resolved molecular fingerprint; QCs - quality controls; PCA - principal component analysis.

The PCA of the EMFs recorded for the QCs is used to analyze the technical noise and the efficiency of every preprocessing step. In particular, given that the QCs samples are the exact replicas of the same sample, in the ideal case there should be no clustering according to any experimental parameter, at least in the first principal components addressing the main source of data variability. After Hilbert centering, plotting PC1 against PC2 it is visible that the EMFs cluster according to the measurement day (Figure 3.2c). The characteristics of the exciting pulse are indeed slightly different every day depending on several factors, such as room humidity, optics degradation, thermalization processes or slightly different alignment. The KORA-FF4 measurements have been performed over three months. To compare the EMFs acquired in this large time span, we need to compensate for any day-to-day variation of the exciting pulse and to extend the same correction to the whole time window. This preprocessing step is called *standardization*. As seen in section 2.2.2, the absorbance of each sample is:

$$A = -log(|t(\omega)|^2) \tag{3.1}$$

where t(ω) is the transfer function:

$$t(\omega) = FT(E_s(t))/FT(E_r(t)) \tag{3.2}$$

with $E_s(t)$ and $E_r(t)$ the electric fields of a generic sample and the respective reference. Because of the day-to-day technical variability, we can assume that two water references $E_{r1}(t)$ and $E_{r2}(t)$ measured in different days are not the same. Therefore, the EMFs derived from them would be not comparable being them defined as:

$$EMF_1 = E_{s1}(t) - E_{r1}(t) = t(t)E_{r1}(t) - E_{r1}(t) \tag{3.3}$$

$$EMF_2 = E_{s2}(t) - E_{r2}(t) = t(t)E_{r2}(t) - E_{r2}(t) \tag{3.4}$$

According to equation 3.2, multiplying $E_s(t)$ and $E_r(t)$ by the same arbitrary non-zero complex function $\psi(\omega)$ does not change the transfer function. Therefore, it is possible to make the two EMFs comparable by applying the following filter based on the electric field of one (or more) specified reference $\tilde{E_r}(t)$:

$$\psi_S(\omega) = FT(E_r(t))/FT(\tilde{E_r}(t)) \tag{3.5}$$

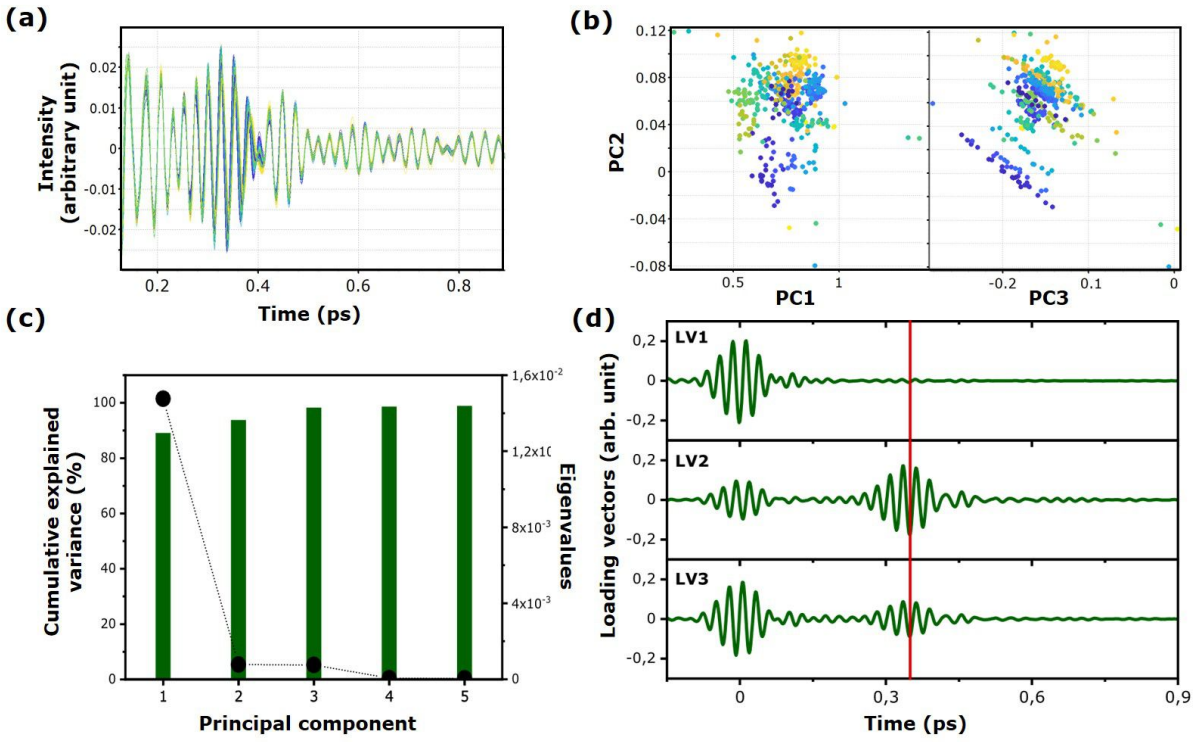

Figure 3.3: FRS preprocessing optimization in the time domain: EMFs standardization. (a) The EMFs of the QCs replica after applying Hilbert centering and standardization show different signatures according to the measurement day around 350 fs. The color gradient highlights the measurement day. (b) The corresponding PCA analysis shows a clustering according to the measurement day along PC2 and PC3. (c) The corresponding eigenvalues and cumulative explained variance of the first five principal components show that PC1 explains about 90% of the total variance, while PC2 and PC3 address about 6% of it together. (d) The loading vectors of the first three principal components show that PC1 depends on the noisy signal of the exciting pulse around zero, while LV2 and LV3 have a strong signature around 350 fs, which reflects the variability shown in panel (a). Acronyms: FRS - field-resolved spectroscopy; EMF - electric-field-resolved molecular fingerprint; QCs - quality controls; PCA - principal component analysis; PC - principal component; LV - loading vector.

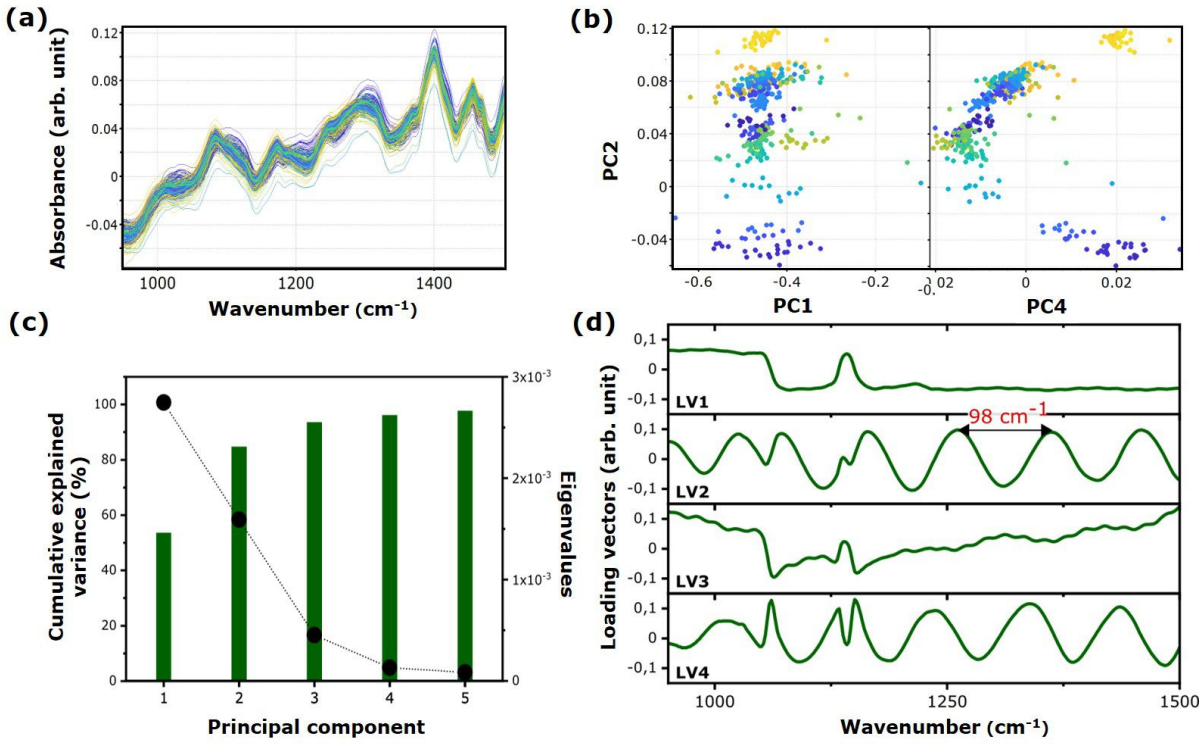

Figure 3.4: FRS preprocessing optimization in the frequency domain: EMFs standardization. (a) The absorption spectra corresponding to the EMFs of QCs shown in Figure3.3a features a high variability. (b) The PCA analysis of the absorption spectra highlights that PC2 and PC4 feature a clustering according to the measurement day. (c) The eigenvalues and cumulative explained variance of the first five principal components show that PC1 addresses about 58% of the total variability, while PC2 and PC4 address about 26% and 2% of it respectively. (d) The loading vectors of the first four principal components show that PC1 is associated with the broad excitation pulse residual, as in the time domain, while LV2 and LV4 feature an interference pattern with a period of 98 cm^{-1} associated with the 350 fs signal seen in Figure3.3. Acronyms: FRS - field-resolved spectroscopy; EMF - electric-field-resolved molecular fingerprint; QCs - quality controls; PCA - principal component analysis; PC - principal component; LV - loading vector.

Standardization completely eliminates the day-to-day dependence in the EMFs of the water references. The PCA of the EMFs of QCs shows that, after standardization, PC1 accounts for the difference in the intensity of the main pulses, as can be seen in the first loading vectors (LV1) both in the EMFs (Figure 3.3d) and in the associated absorption spectra (Figure 3.4d). However, in the time domain PC2 and PC3 still shows clustering according to the measurement day (Figure 3.3b). This is also evident along PC2 and PC4 of the respective absorption spectra (Figure 3.4b), for which the corresponding LVs feature an interference pattern with a period of 98 cm^{-1} (Figure 3.4d). This pattern corresponds to the signature at 350 fs in the time domain (Figure 3.3a, d). A similar interference pattern arises at 1.9 ps, exactly 350 fs after the first back reflection of the exciting pulse at the EOS crystal. Despite the day-to-day dependence is reduced, it still accounts for a large part of the data variability (Figure 3.3c and 3.4c). Therefore, it is vital to address its technical origin.

To this end, the EMFs of QCs measured on consecutive days in similar settings are compared. The thickness of the measurement cell has been measured via an FT-IR spectrometer and is

31.89 μm under atmospheric pressure. Under vacuum, the effective thickness D can be retrieved from the definition of group velocity:

$$v_g = D/t = c/n_g \tag{3.6}$$

From the equation above, we can derive the thickness D as $2 * D = (c * dt)/n_{QC}$ by setting dt to 350 fs. The refractive index of QCs (n_{QC}) can be derived from the one of water (1.350) knowing that the EMFs of water and QCs induce a time delay in the main pulse of about 0.5 fs, which returns $n_{QC} = 1.355$. Therefore, the cell thickness under vacuum is 38.72 μm, measured at an ambient pressure of about 950 $mbar$, due to the low pressure on both external sides of the measurement cell. The oscillator, DFG and EOS chambers are operated under vacuum (below 1 $mbar$) to reduce the strong absorption of water. If the effective thickness under vacuum would be constant, this would not introduce any day-to-day dependence. Since the atmospheric pressure is different every day (Figure 3.5b), it influences the effective thickness acting from the top of the cell inducing small changes (dD) from one day to another. Figure3.5a shows the EMFs of QCs measured in two different days with an ambient pressure difference of 0.5 $mbar$. From Equation 3.6 it can be calculated that the delay of about 1 fs in the interference pattern at 350 fs corresponds to a dD of 0.22 μm (0.6% of the cell thickness). For the extreme cases of the very first and the very last QCs measured, performed with an ambient pressure difference of about 14 $mbar$, the delay is about 3.6 fs and corresponds to a dD of 0.79 μm, about 2% of the cell thickness.

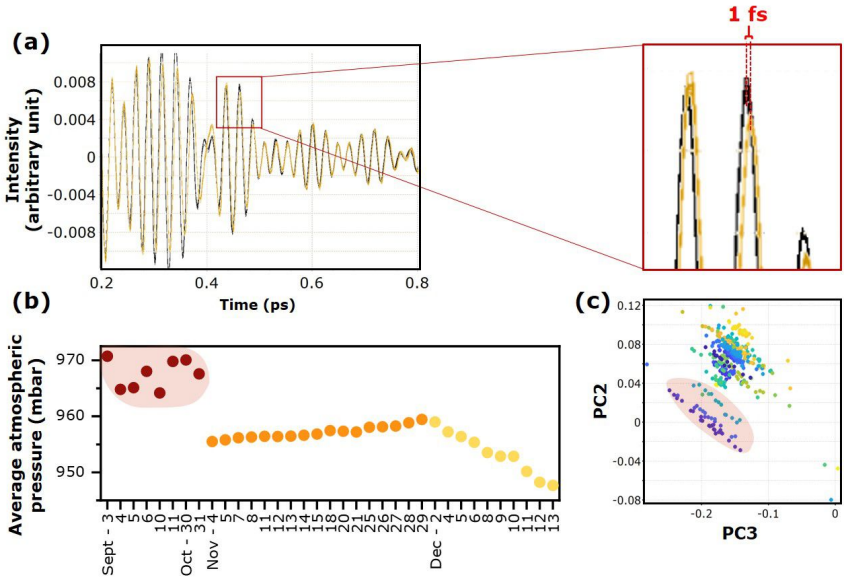

Figure 3.5: Day-to-day dependence of FRS measurements: effect of the ambient pressure on the measurement cell thickness. (a) The EMFs of QCs measured on consecutive days, colored by measurement day, show a delay of about 1 fs around 350 fs (inset). (b) The ambient pressure for each measurement day of the KORA-FF4 measurement changes from day to day (colored by month). (c) The PC1/PC2 plot for the EMFs of the QCs shows that the clustering is connected with the ambient pressure variation (colored by measurement day). The red shaded area in panels (c) and (b) refer to the same month. Acronyms: FRS - field-resolved spectroscopy; EMF - electric-field-resolved molecular fingerprint; QCs - quality controls; PC - principal component.

Ultimately, it can be concluded that the day-to-day dependence arises from an internal reflection of the exciting pulse inside the measurement cell because of the thickness variations of the last in function of the external pressure. However, this does not explain why standardization is not fully efficient on biological samples. This is due to their slightly different refractive indexes between the water references and the actual samples. Fortunately, because of their similar chemical composition, the human blood serum replica QCs and the actual human blood plasma samples have very close refractive indexes, which turns vital for the implementation of a preprocessing step able to reduce the day-to-day dependence in the EMFs of the samples. This step is called *interference correction*. Each sample is associated with the closest QC measured to build the following filter:

$$\psi_{IC}(\omega) = \psi_S(\omega) * (t_s(\omega)t_{r(QC)}(\omega))/(t_{QC}(\omega)t_r(\omega)) \tag{3.7}$$

where $t_{s,r}$ are the transfer functions of the sample we apply the correction to and its reference, $t_{QC,r(QC)}$ are the ones associated to the closest QC and $\psi_S(\omega)$ is the filter used for standardization.

Figure 3.6c shows the effect of different preprocessing procedures on the technical noise. Standardization lowers the technical noise of the main pulse to the minimum, but makes longer times noisier. Interference and *echo correction* compensate the noise at 350 fs and 1.5 ps selectively, the last originating from the back reflection at the EOS crystal. Applying echo correction after the other preprocessing steps acts specifically at 1.5 ps. However, if applied before, it helps standardization and leads to a strong reduction of the technical noise at longer times. Therefore, it can be concluded that the preprocessing yielding the lowest standard deviation of the EMFs of QCs is based on the following steps applied in the specified order:

1. Hilbert centering (HC)
2. Echo correction (EC)
3. Interference correction (IC)
4. Standardization (ST)

Thanks to this combination of preprocessing steps the day-to-day dependence of the EMFs of QCs is extensively reduced (Figure3.6a, b).

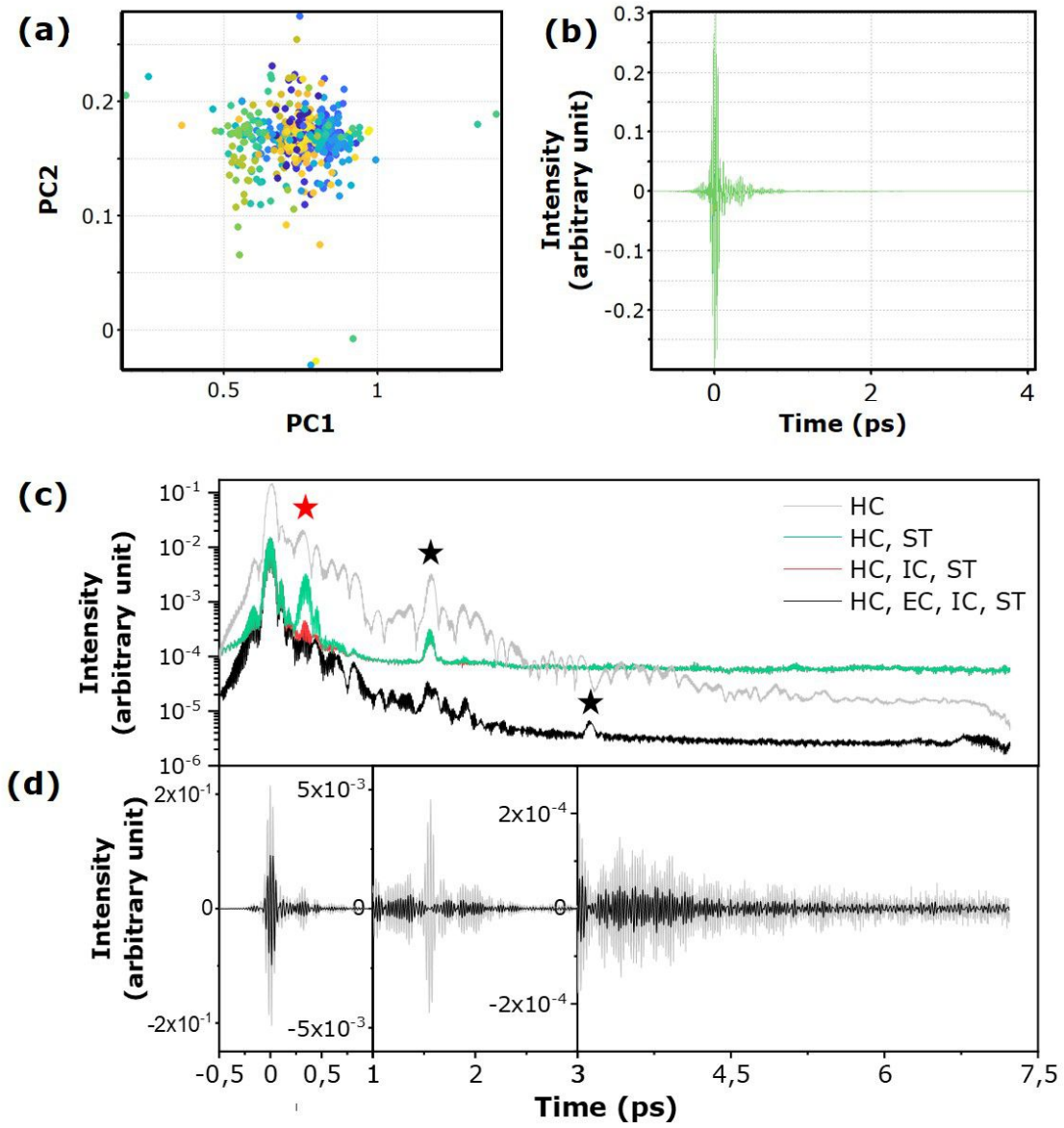

Figure 3.6: Optimal preprocessing for the KORA-FF4 EMFs in the time domain. (a) The PC1/PC2 plot of (b) the EMFs of all QCs after applying the optimized preprocessing protocol show no clustering according to the measurement day, highlighting that the preprocessing proposed extensively reduces the day-to-day dependence, giving identical EMFs for the replicas QCs samples (colored by measurement day). (c) The standard deviation of the EMFs of all QCs is shown for different preprocessing highlighting that the optimized preprocessing (black line) minimizes the technical noise. (d) The comparison of the same EMF of a QC after applying Hilbert centering only (grey) and applying the optimized preprocessing protocol (black) shows the extensive reduction of the technical noise. Acronyms: EMF - electric-field-resolved molecular fingerprint; PC - principal component; QCs - quality controls; HT - Hilbert centering; EC - echo correction; IC - interference correction; ST - standardization; EOS - electro-optic sampling. Red star - interference pattern; black stars - back reflection at the EOS crystal.

3.1.3 Comparison of biological and technical noise of IR fingerprints

The optimal preprocessing protocols found for the FT-IR and FRS data of QCs are applied on the KORA-FF4 samples to compare the technical noise characterized in the previous sections with the biological one, namely the spectral variation of the samples due to the biological variability between individuals. The aim in this section is to compare the biological-to-technical noise ratio obtained via the newly developed FRS spectroscopy with a state-of-the-art FT-IR spectrometer.

The PCA of QCs and samples show a day-to-day dependence only along the PC1 after applying only the Hilbert centering preprocessing step (Figure 3.7a). This highlights a higher impact of the day-to-day dependence compared to the difference between QCs (black shaded area) and samples, affecting PC2. The optimal preprocessing removes the influence of the measurement day and efficiently reduces the standard deviation of human blood plasma EMFs (Figure 3.7b).

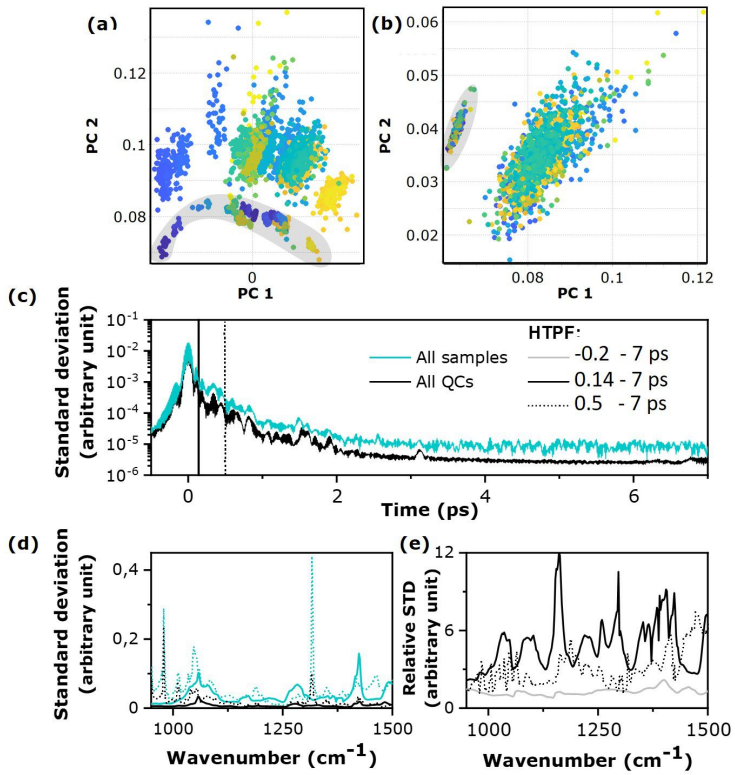

Figure 3.7: Biological-to-technical noise ratio of EMFs in the time and frequency domain. (a) The PC1/PC2 plot for the EMFs of samples and QCs (black shaded area) with Hilbert centering only and (b) applying the optimized preprocessing protocol shows that the preprocessing reduces the day-to-day dependence of the EMFs. (c) The standard deviation of the EMFs of samples and QCs in the time and (d) the frequency domain for the HTPFs highlighted in panel (c). (e) The relative standard deviation calculated as the ratio between the standard deviation of the EMFs of samples and the one of QCs shows the biological-to-technical noise ratio for the HTPFs shown in panel (c), highlighting that the ratio is maximized for the HTPF going from 0.14 to 7 *ps*. Acronyms: EMF - electric-field-resolved molecular fingerprint; PC - principal component; QCs - quality controls; HTPF - high temporal pass filter; STD - standard deviation.

Figure 3.7c shows that the biological variability encoded in the EMFs of the biosamples is higher than the technical noise in the whole time window starting from 140 fs. The residual signal of the exciting pulse is a major source of noise that can be removed with an HTPF. The absorption spectrum can be calculated from EMFs via Equation 3.1 as $-log(|t(\omega)|^2)$, with $t(\omega)$ the Fourier transform of the corresponding time trace. To reduce the noise coming from the exciting pulse, an HTPF is applied and the Fourier transform is defined as $t_{HTPF}(\omega)$. The $-log(|t_{HTPF}(\omega)|^2)$ is not an absorbance, but it allows a closer comparison with FT-IR spectra and will be considered as the EMFs' signal in the frequency domain. The standard deviations of both samples and QCs in the frequency domain for different HTPFs increase for longer initial times (e.g. going from the temporal window 0.14 - 7 ps to 0.5 - 7 ps). The highest biological-to-technical noise ratio is associated with an HTPF from 0.14 to 7 ps which will therefore be employed in all the following analyses (Figure 3.7d, e).

The FT-IR analysis of QCs highlights that the technical noise is higher for the Amide bands of proteins (1250 - 1750 cm^{-1}; Figure 3.8b, black line). The biological noise is higher for the signatures of proteins and lipids (1750 - 3000 cm^{-1}; Figure 3.8b, light purple line). The FT-IR analysis of KORA-FF4 blood plasma samples highlights that the biological-to-technical noise ratio is smaller for proteins, while lipids are the main source of the biological variability between individuals (see section 3.2.2). In the spectral range between 1000 and 1250 cm^{-1}, the biological-to-technical noise ratio calculated for FT-IR is higher than for the Amide bands as well as compared to the FRS data in the frequency domain. Between 1250 and 1500 cm^{-1}, where the ratio calculated for FT-IR is smaller, the two techniques reach similar values (Figure 3.8c).

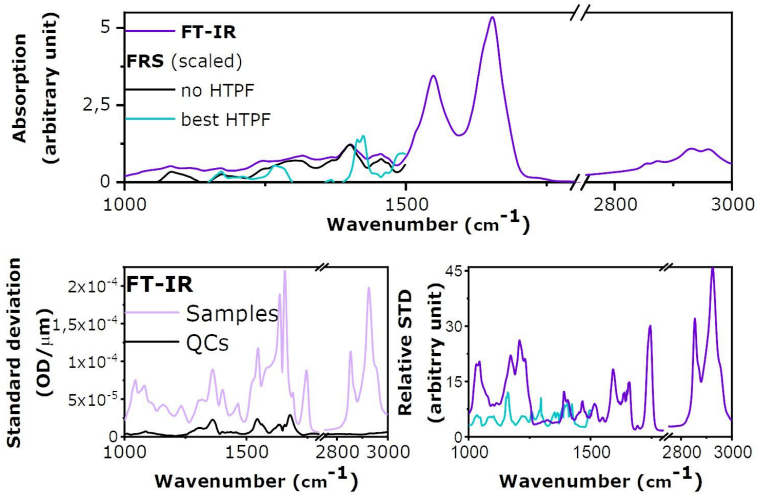

Figure 3.8: Biological-to-technical noise ratio of FT-IR and FRS in the frequency domain. (a) The average FT-IR spectra of all KORA-FF4 samples are compared with the FRS spectra derived with and without a HTPF (0.14 - 7 ps). (b) The standard deviation of FT-IR spectra of samples and QCs show a higher technical noise from the amide bands of proteins and a higher biological noise from proteins and lipids. (c) The biological-to-technical noise ratio is compared for FT-IR and FRS in frequency domain showing that the ratio is comparable between 1250 and 1500 cm^{-1} and higher for FT-IR at lower frequencies. Acronyms: FRS - field-resolved spectroscopy; HTPF - high temporal pass filter; QCs - quality controls; STD - standard deviation.

In conclusion, the spectra and EMFs of QCs have been first evaluated for both FT-IR and FRS for the identification of the optimal preprocessing able to minimize the technical noise associated with the measurement of liquid blood-based biosamples. The optimized preprocessing found for each technique has then been applied on the KORA-FF4 blood plasma samples which have been used to identify the biological noise due to the biological variability between individuals. The biological-to-technical noise ratio highlights a larger biological variability between individuals associated with the spectral signatures characteristic of lipids. The comparison of the biological-to-technical noise ratio derived for the newly developed FRS spectroscopy with the one obtained via a state-of-the-art FT-IR spectrometer highlights that the performances are comparable between 1250 and 1500 cm^{-1}, but are still smaller for FRS at shorter wavenumbers. Further improvements on the FRS setup are currently being implemented to boost its performance.

3.2 Between-person spectral variability among healthy individuals

Infrared fingerprinting is a fast and cost-efficient way of recording a snapshot of the chemical composition of a sample. The standard deviation of the IR fingerprints measures the biological variability of human blood biofluids between individuals. Despite the last is well known to have a strong impact on FT-IR spectra [58, 60], there are no comprehensive studies that try to tackle the source of this variability in a general large population. The FT-IR fingerprints and the FRS time traces of the KORA-FF4 cross-sectional population-based cohort are perfectly suited for this purpose. Before tackling this issue, the KORA-FF4 cohort is shortly introduced.

3.2.1 KORA-FF4 cross-sectional population-based cohort

In this section, the KORA-FF4 cross-sectional population-based cohort is introduced. Table 3.1 shows the number of individuals positive for each known intermediate and endpoint medical condition (Figure 3.9a) as well as of the healthiest individuals, namely the ones non-symptomatic (NSP) to the known medical conditions and with normal glucose tolerance (NGT), identified as NSP/NGT.

In particular, the known medical conditions for this cohort are: prediabetes, type II diabetes, heart attack (individuals who had an episode of myocardial infarction), hypertension, high blood lipids, chronic obstructive pulmonary disease (COPD), asthma, as well as individuals who had cancer (former or ex-cancer cohort) or a stroke. Being a cross-sectional population-based cohort, the number of cases reflects the probability of developing that disease in the represented population. Hypertension and high blood lipids affect about 50% of the population, with more than a thousand cases each. Fewer individuals are positive to the other medical conditions, ranging from 60 cases that had a stroke to about 350 individuals with prediabetes. The number of cases is important for this type of study as it influences the outcome of the binary classifications (see section 4.2.3).

The common parameters evaluated in this study are easily accessible factors, such as gender, age, smoking status, alcohol consumption and body mass index (BMI) which is defined as the ratio between a person´s weight and the square of his height. Another less accessible parameter

considered is the C-reactive protein (CRP) concentration, indicative of many physiological changes and commonly used as an index for the inflammation level [89]. The average values reported in Table 3.1 reveal that the healthiest individuals (NSP/NGT) have the lowest average age, BMI and CRP concentration, while there are no major differences with the other cohorts for the ratio between smokers and non-smokers and the average daily alcohol consumption.

KORA-FF4 cohort							
Cohort	n. cases	M/F	Age	CRP (mg/L)	BMI (kg/m^2)	Smok / not smok	Alcohol (g/day)
All	2074	1.0	60.2 ± 12.3	2.4 ± 4.4	27.8 ± 4.9	0.2	14.7 ± 20
NSP/NGT	394	0.7	51.6 ± 10	1.5 ± 2.6	25.4 ± 3.7	0.3	13.8 ± 17
Prediabetes	360	1.4	65.4 ± 11.1	3.1 ± 4.9	29.6 ± 4.7	0.2	17.8 ± 23.2
Diabetes	288	1.5	69.4 ± 10.1	3.5 ± 5	31.1 ± 5.4	0.1	15.9± 24.3
Heart attack	69	3.1	71.1 ± 9.4	3.1 ± 4.8	30.3 ± 5.5	0.1	13.7 ± 20.8
Hypertension	1035	1.1	64.5 ± 11.2	2.8 ± 4.5	29.3 ± 5.2	0.1	15.1 ± 21.1
High lipids	1033	1.6	62.1 ± 11.5	2.6 ± 5.2	28.3 ± 4.8	0.2	15.8 ± 22.1
Stroke	54	1.3	72 ± 9.1	4 ± 5.9	29.2 ± 4.2	0.1	16.2 ± 23.7
COPD	150	0.8	64.8 ± 11.2	4.2 ± 8.1	30 ± 6.1	0.2	12.5 ± 17.7
Ex-cancer	229	1.1	67.1 ± 11.5	2.7 ± 4.6	27.7 ± 4.4	0.1	15.2 ± 19.1
Asthma	182	0.7	60.1 ± 12	2.6 ± 3.2	28.2 ± 5.6	0.2	11.9 ± 17.3

Table 3.1: Description of the KORA-FF4 cohort. The table shows the number of individuals positive for each known common medical condition and of the healthiest non-symptomatic individuals with normal glucose tolerance (NSP/NGT). For each cohort, the average values of age, CRP, BMI and daily alcohol consumption are reported together with the ratio between the number of male and female individuals and the one between active smokers and non-active smokers (former and never smokers). Since KORA-FF4 is a cross-sectional population-based cohort, the number of cases show the probability of developing each medical condition in the given population. The NSP/NGT individuals have the lowest average age, BMI and CRP compared to the other cohorts. Acronyms: M/F - males-to-females ratio; CRP - C-reactive protein; BMI - body mass index; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; COPD - chronic obstructive pulmonary disease.

3.2.2 FT-IR fingerprints of common parameters in KORA-FF4

This section focuses on the analysis of the FT-IR fingerprints to identify the main common parameters responsible for the between-person spectral variability among the healthiest individuals, namely the NSP/NGT cohort. To that end, both unsupervised machine learning algorithms and SVM binary classifications are applied. The common parameters considered are shown in Table 3.1. To analyze the effect of each parameter, the NSP/NGT individuals are grouped into case/control cohorts according to the definitions reported in Table 3.2.

Parameter	Control cohort	Case cohort
Gender	females	males
Age	<55 years old	> 54 years old
BMI	< 25 kg/m^2 - underweight and normal weight	> 24.99 kg/m^2 - preobese and obese
Smoking status	non-active smokers	active smokers
Alcohol consumption	0-20 g/day	> 20 g/day
Inflammation	[CRP] < 5 mg/L - low	[CRP] > 4.99 mg/L - high

Table 3.2: KORA-FF4 case/control cohorts for each common parameter among healthy individuals. The known common parameters listed in Table 3.1 are here analyzed in case-control studies. The thresholds used to identify these cohorts, which are based on different risk categories, are here listed for each parameter. Acronyms: BMI - body mass index; CRP - C-reactive protein.

The between-person spectral variability, namely the standard deviations of the FT-IR spectra of a selected cohort of individuals, have been computed for each medical condition (Figure 3.9b, grey bars) as well as for each respective sub-cohort defined according to Table 3.2 (Figure 3.9b, colored bars). The NSP/NGT individuals are expected to have the smallest between-person variability, in agreement with what observed in the FT-IR data presented here. Prediabetes, an intermediate condition, induces a higher biological variability between individuals, slightly smaller than the endpoint diseases. All medical conditions reach comparable standard deviations, most probably because each individual might be affected by the other known conditions as well. Indeed, most symptomatic cases have at least one or two comorbidities and only too few individuals have none, making it difficult to disentangle the effect of every single condition on the between-person spectral variability. Among the healthiest individuals, the spectral variability of males is larger compared to females. The same is true for smokers, preobese and obese individuals, people with high daily alcohol consumption and individuals with high inflammation compared to their respective controls (Table 3.2). These trends hold for symptomatic individuals, with a few exceptions (Figure 3.9b, *). For example, diabetes introduces a larger between-person spectral variability among individuals younger than 55 years compared to the older ones and, together with prediabetes, heart attack and stroke, to the preobese and obese individuals compared with the ones with normal weight.

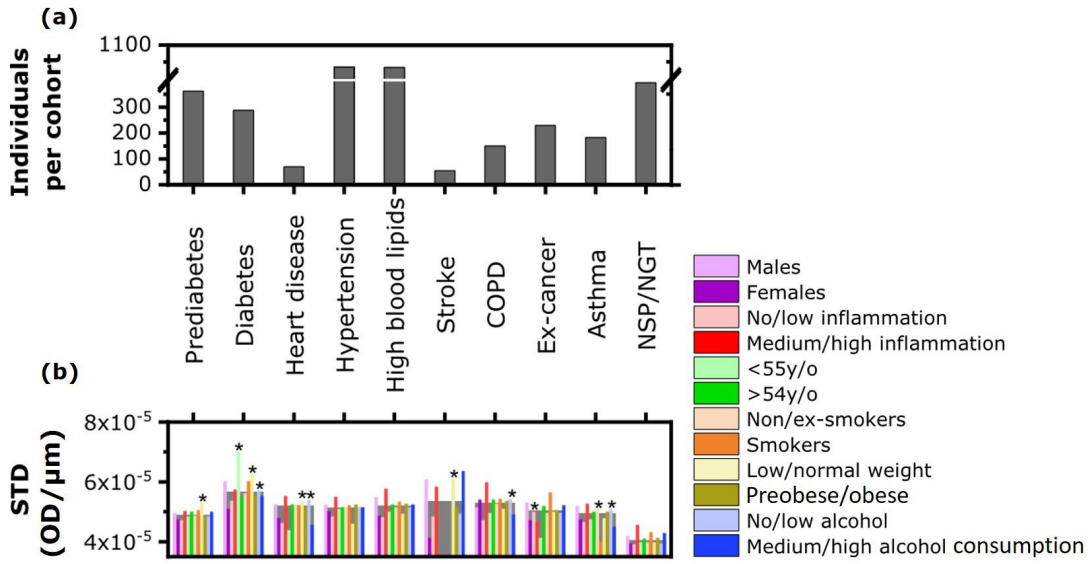

Figure 3.9: Number of individuals for each medical condition in KORA-FF4 and corresponding FT-IR between-person spectral variability. (a) The number of individuals positive for each known medical condition is reported for the KORA-FF4 population (Table 3.1). (b) The between-person spectral variability, namely the total standard deviation of the spectra, is reported for each cohort in panel (a) (black shaded bars) and the respective sub-cohorts based on the known common parameters listed in Table 3.2 (colored bars: legend). Acronyms: COPD - chronic obstructive pulmonary disease; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; y/o - years old.

Analysis via SVM binary classifications

This section wants to address to what extend each known common parameter affects the FT-IR fingerprints of the human blood plasma of NSP/NGT individuals. In particular, this is addressed by performing binary classifications of the case-control cohorts defined in Table 3.2. Each binary classification is performed via the supervised machine learning algorithm SVM in 10-fold cross-validation (repeated 10 times).

The efficiency of each classification, measured with the AUC of the ROC curves, addresses the impact of each parameter on all spectral features. In particular, higher AUCs are found for gender and inflammation (> 80%), followed by smoking status, age and alcohol consumption (70 - 80 %), while BMI has a relatively low AUC of 66 % (Figure 3.10a). A strong impact from physical activity is expected according to what has been reported in the literature [90, 91]. Surprisingly, the classification of physically active against non-active individuals results in no separation (AUC = 50%), probably because of the qualitative and strongly biased nature of this self-reported parameter.

Figure 3.10b shows the differential fingerprint associated with each parameter, calculated as the difference between the average FT-IR spectrum of the cases and the average spectrum of the controls. The SVM coefficients, which unveil the features responsible for each binary classification, are in good agreement with the differential fingerprints. The only exceptions are inflammation, smoking status and alcohol consumption for which the SVM coefficients are too noisy because of the low number of cases (< 150 individuals, see section 4.2.3).

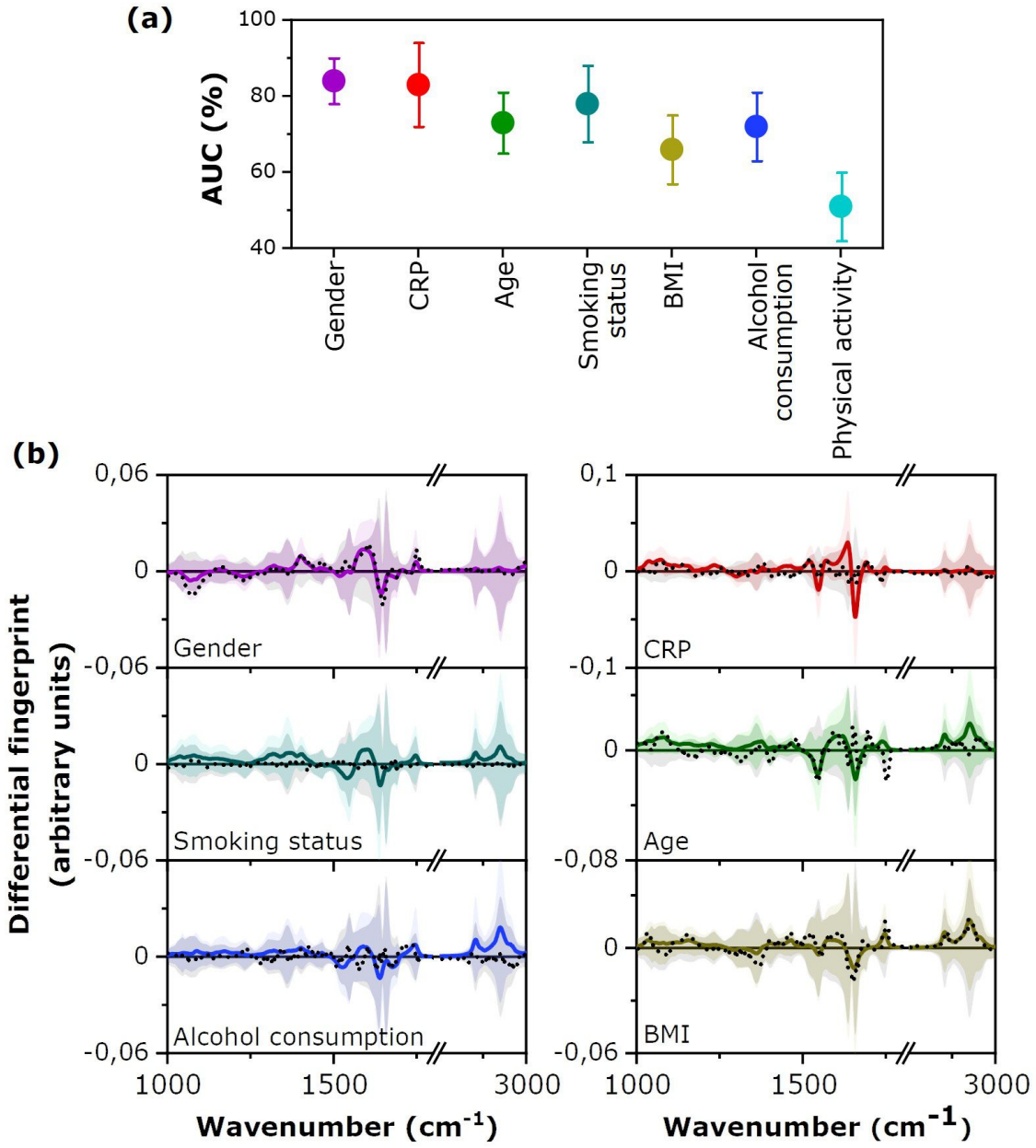

Figure 3.10: SVM binary classification of common parameters on the FT-IR spectra of healthy individuals in KORA-FF4. (a) The AUCs obtained via SVM binary classification for the known common parameters among NSP/NGT individuals (Table 3.2) unveils that gender and inflammation have the strongest impact on the fingerprints, followed by age, smoking status and alcohol consumption, while BMI and physical activity have low or no impact. (b) The differential fingerprint of each parameter is comparable to the SVM coefficients (black dotted lines), which unveil the spectral signatures responsible for the binary classification, with the exceptions of inflammation, smoking status and alcohol consumption which count few cases leading to noisy SVM coefficients. Shaded areas: standard deviations of each cohort in the differential fingerprint (colored: case cohort; black: control cohort). Acronyms: SVM - support vector machine; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; BMI - body mass index; CRP - C-reactive protein.

From this analysis, it is evident that gender and inflammation have the strongest impact on the FT-IR fingerprints of healthy. In particular, the most relevant features associated with gender arise mostly at low frequency, between 1000 and 1250 cm^{-1}, with some contributions from the amide bands. The differential fingerprint associated with CRP has a strong protein signature characteristic of the altered albumin-to-globulin ratio (AGR) in the Amide I and II bands (1500 – 1800 cm^{-1}), characteristic of inflammation [61, 92]. Age, smoking status and daily alcohol consumption have also an important impact on the IR fingerprints of healthy, all with a relevant signature in the protein and lipid spectral region (the last between 1800 and 3000 cm^{-1}, see Table 2.1). The parameter with the smallest impact is BMI, which shows similar spectral contributions as age and CRP.

Analysis via unsupervised algorithms

Figure 3.9b shows that, among NSP/NGT individuals, medium-to-high inflammation is associated with the largest between-person spectral variability and the SVM supervised analysis shows that inflammation, together with gender, has the strongest impact on the FT-IR fingerprints of healthy. The classification efficiencies are informative for the impact of each parameter on all spectral features but are not sufficient to identify their impact on the between-person spectral variability specifically. A deeper analysis based on unsupervised methods is here proposed to identify clusters of similar spectral fingerprint, analyze how the common parameters are distributed between the clusters and understand which of them has the strongest impact in the clustering and, therefore, on the IR signatures. All the analyses reported here are performed on the PCs encoding 99% of the total variance.

The optimal number of clusters is first obtained via the elbow method and it is found to be three. This is indeed the number of clusters leading to a drastic change in the slope of the within-cluster sum of squares (WCSS) (Figure 3.11c). K-means, agglomerative clustering and Gaussian mixture model (GMM) are applied and compared via their silhouette scores. K-means and GMM return comparable scores (Figure 3.11a). In particular, GMM performs better with covariance type full compared to setting the covariance type as spherical (Figure 3.11b). Both K-mean and GMM find similar clusters with a large separation along PC1 and PC2 (Figure 3.11e-d respectively).

The distribution of the common parameters in each cluster highlights their influence on the clustering outcomes. From Table 3.3 and Figure 3.11f, it can be seen that age and inflammation play an important role in defining the clusters. In particular, most of the individuals with medium-to-high inflammation end up in cluster 1 and most of the individuals older than 55 years, especially if they have high risk factors (smokers, high alcohol consumption, high BMI), end up in cluster 3. Therefore, most of the individuals with low risk factors, age and inflammation are in cluster 2, separated from cluster 1 along PC2 and from cluster 3 along PC1.

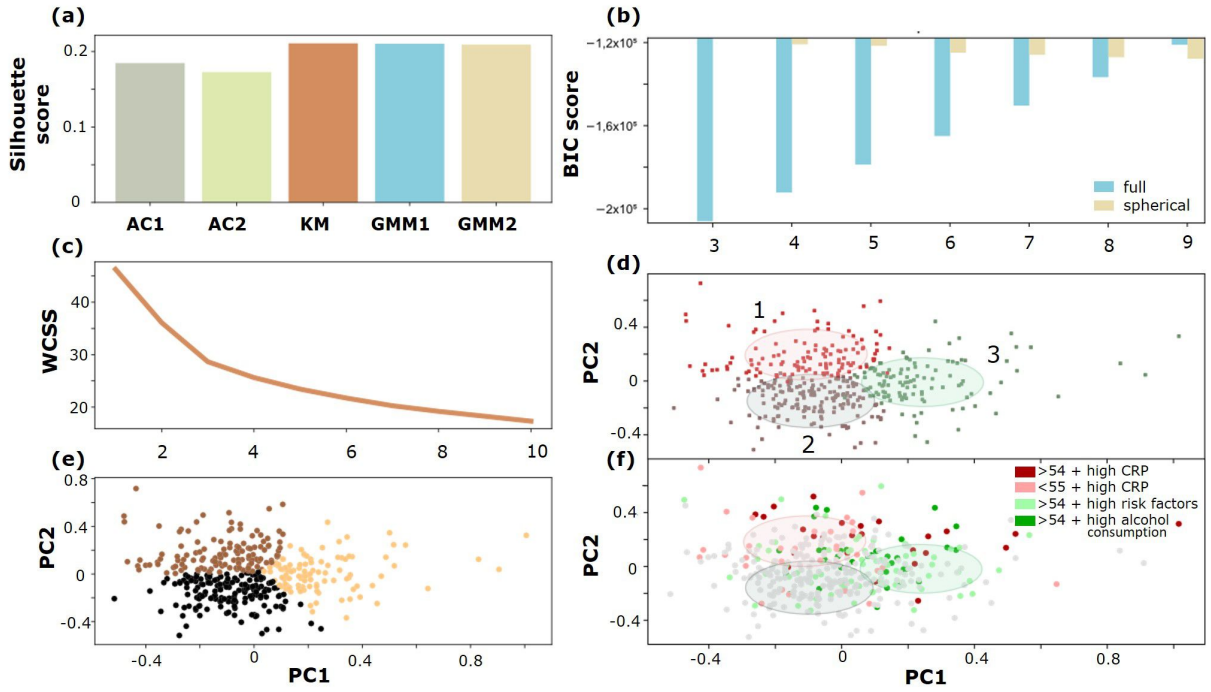

Figure 3.11: Unsupervised clustering analysis of the FT-IR spectra of healthy individuals in KORA-FF4. (a) The silhouette scores are reported for the unsupervised algorithms applied on the PCs explaining 99% of the total variability calculated for the NSP/NGT cohort. The graph shows that K-means and GMM have the best performances. (b) The BIC scores found for GMM clustering with different number of components and covariance types shows the highest performance for GMM with covariance full for 3 clusters. (c) The elbow method reporting WCSS against the number of clusters shows 3 clusters as the optimal number. (d) The PC1/PC2 plot colored according to the clusters identified via GMM (covariance type full) and (e) K-means are compared with (f) the PC1/PC2 plot colored according to selected combinations of common parameters (colored in grey are the individuals that do not correspond to any of the specified groups). The comparison highlights that K-means and GMM return similar clusters. Moreover, it highlights that the datapoins corresponding to individuals with high inflammation (high CRP) tend to cluster together, as well as for the data corresponding to individuals older than 55 years, especially for high risk factors (smokers, high alcohol consumption or high BMI). Acronyms: PC - principal component; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; BIC - Bayesian information criterion; WCSS - within-cluster sum of squares; CRP - C-reactive protein; BMI - body mass index; AC - Agglomerative clustering (1 - affinity = 'euclidean', linkage = 'ward'; 2 - affinity = 'cosine', linkage = 'average'); KM - K-means; GMM - Gaussian Mixture Model (1 - covariance type = 'full'; 2 - covariance type = 'spherical').

Cluster	young with high CRP concentration	old with high CRP concentration	old with high risk factors
1	45.2%	42.3%	32.3%
2	25.8%	23.1%	22.6%
3	29.0%	32.1%	45.2%

Table 3.3: GMM clusters found for the FT-IR fingerprints of healthy individuals in KORA-FF4. Age and inflammation are the most unequally distributed parameters among the three clusters. The percentage of individuals with high inflammation and older than 55 years are reported. Most of the individuals with high inflammation are in cluster 1, while most of the individuals older than 55 years and with high risk factors (smokers, high alcohol consumption or high BMI) are in cluster 3. Therefore, most of the individuals with low risk factors, age and inflammation are in cluster 2. Acronyms: GMM - Gaussian Mixture Model; CRP - C-reactive protein; BMI - body mass index.

The unsupervised clustering analysis highlight that the first two PCs address the origin of the between-person variability. Among all the common factors analyzed, age and inflammation seem to have the strongest impact. Therefore, despite the high AUC associated with gender, this parameter does not influence the features connected specifically to the between-person variability. However, many unknown parameters can play a crucial role in the clustering (e.g. hormonal status, cholesterol level, etc.). The analysis performed here considers only the most common and easily accessible factors. Despite CRP concentrations are not immediately available, this analysis highlights the huge impact of inflammation on the spectral variability between individuals. Ignoring it would possibly introduce biases to the analysis of the IR fingerprints (see section 4.2.1). To address the spectral features affected by the between-person variability, it is necessary to further investigate the first two PC components.

Principal component analysis

The variability explained by each PC abruptly decreases for higher components, with PC1 and PC2 addressing for about 70% of the total variance of the FT-IR spectra of NSP/NGT individuals (Figure 3.12a). Unsupervised clustering algorithms highlight that considering PC1 and PC2 is enough to address the origin of the between-person variability among healthy individuals. Therefore, these PCs are used to identify the features encoding for it. The spectral assignments to the different classes of biomolecules refer to Table 2.1.

LV1 and LV2 show that PC1 addresses the variance of the lipid signature (1800 – 3000 cm^{-1}; Figure 3.12b), which accounts for about 45% of the total variability, and PC2 addresses the variance of the AGR signature seen for inflammation (Figure 3.12b). Figure 3.12c shows that PC1 correlates with age, which therefore affects the variability of lipids among healthy individuals, in agreement with the previously acknowledged up-regulation of lipids concentration with age [93, 94]. PC2, instead, correlates with age and CRP concentrations (Figure 3.12c), showing that these parameters influence the AGR protein signature in healthy individuals. Other factors such as BMI and smoking status have small correlations with PC1 and PC4, but the p-values are close to 0.01, taken as threshold for meaningful correlations.

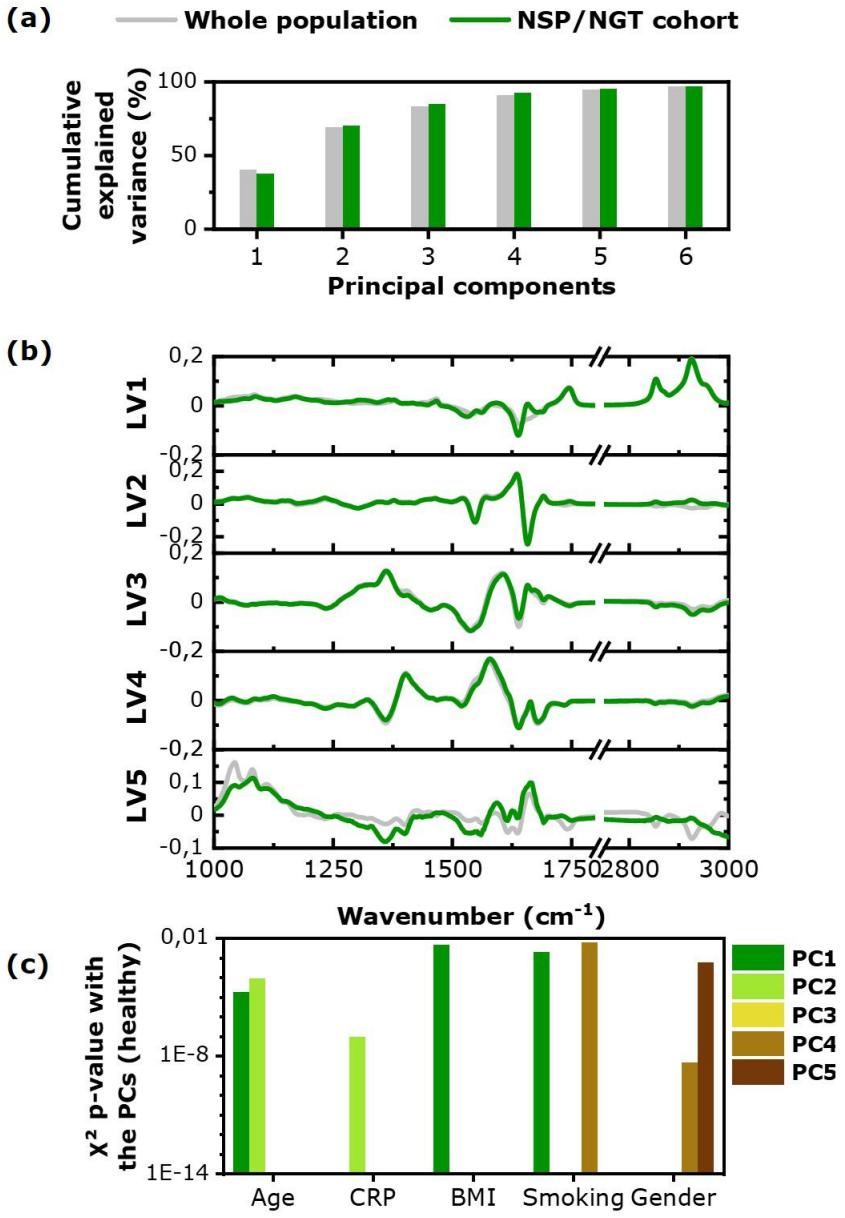

Figure 3.12: PCA analysis of the FT-IR spectra of healthy individuals in KORA-FF4. (a) The cumulative explained variance of the first 5 PCs and (b) the corresponding loading vectors are compared for the NSP/NGT cohort and the whole population. The comparison highlights that the features encoding for the spectral variability is analogous for symptomatic and non-symptomatic individuals. LV1 addresses the lipid signature (Table 2.1), LV2 the protein AGR signature associated with inflammation and LV5 the signatures from carbohydrates and protein glycosylation (Figure 3.10b). The spectral nature of the other components is not unambiguously determined. (c) The χ^2 derived p-values below 0.01 show the significant correlations between the common parameters and the first five PCs among NSP/NGT individuals. Smaller values imply stronger correlations. PC1 and PC2 have stronger correlations with age and CRP concentration, while gender correlates with PC4 and with PC5. BMI and smoking status have small correlations with the PCs. Acronyms: PCA - principal component analysis; PC - principal component; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; LV - leading vector; AGR - albumin-globulin ratio; CRP - C-reactive protein; BMI - body mass index.

Contributions in the same spectral region covered by FRS (1000 - 1500 cm^{-1}) arise only from LV3. Because of the different spectral coverage, the spectral features and, therefore, the associated between-person spectral variability addressed by the two techniques have a different molecular origin. For example, between 1000 and 1250 cm^{-1}, the spectral signature are mostly coming from carbohydrates and proteins glycosylations, as described in Table 2.1. PC5 has a strong signature in this spectral region (Figure 3.12d), showing that these classes of molecules are responsible only for about 2% of the between-person spectral variability in the whole spectral coverage of FT-IR. Gender correlate with PC5 and are therefore relevant for the between-person variability for the FRS data (Figure 3.12a).

In conclusion, the correlations of the first two PCs with age and inflammation, together with the high SVM classification efficiencies found for these parameters, highlight that these are the main factors responsible for the large between-person variability among healthy, respectively associated with lipids and proteins variability. These conclusions agree with the ones obtained via unsupervised clustering (Figure 3.12d).

Discussion

In summary, both supervised and unsupervised methods unveil that age and inflammation are the main sources of the between-person spectral variability in FT-IR spectra, with age connected with the highest source of variability between healthy individuals, namely lipids (1750 - 3000 cm^{-1}), as well as with the second-highest source of variability, the proteins (1250 - 1750 cm^{-1}), and inflammation affecting only the second. Gender is shown to be relevant for the biological variability associated with carbohydrates and protein glycosylations (1000 - 1250 cm^{-1}), accounting for a small percentage of the total variability in full spectra, but main actors in the spectral coverage of FRS. It is moreover noticeable that both cumulative explained variance and loading vectors are almost identical for NSP/NGT individuals and the whole population (Figure 3.9a-b). Therefore, it can be concluded that the source of the larger between-person variability found in intermediate and endpoint diseases (Figure 3.9b) are the same features, namely the same biomolecules, responsible for the biological variability among the healthiest individuals. Knowing that the main sources of variability among healthy are age and inflammation, it does not surprise that symptomatic individuals, who have a higher average age and inflammation level (Table 3.1), have a higher between-person spectral variability. These findings underline how understanding the origin of the between-person spectral variability among healthy individuals is fundamental before applying FT-IR spectroscopy for disease diagnosis (see chapter 4).

The impact of age, inflammation and, partially, BMI on the between-person variability might be due to the connection existing between these factors, such as the one between CRP and obesity [95, 96], between aging and weight loss as well as between aging and inflammation pathways [97]. Moreover, low-grade chronic inflammation is associated with aging and is therefore called “inflammaging”, which potentially explains why most of the age-related diseases share an inflammatory pathogenesis [98]. The analysis of serum samples of 590 healthy individuals who participated in the KORA S4 and F4 studies to identify age-associated metabolites [99] support the theory according to which aging is linked with altered lipid metabolism [100]. This agrees with the connection between age and the spectral signatures of lipids. The connection between age and inflammation might explain the impact of age on the AGR signature.

Unsupervised clustering has been used also to define specific thresholds for age, CRP and BMI to group the spectra in new clusters based on these values. However, the so-obtained clusters are not well separate from the ones identified via unsupervised methods. There could be multiple reasons for that. For example, there could be other unknown parameters with similar importance that need to be considered when defining the clusters. A more interesting explanation is that age is not a robust parameter because it does not represent the "effective age" of a person; different individuals are indeed found to go through aging at different rates and, potentially, via different mechanisms [97]. Therefore, each individual of a certain age could potentially find him/herself at a different stage of the aging process and have, therefore, a different blood composition compared to other individuals of the same age. Moreover, the association of aging, inflammation and BMI might have a combined effect difficult to define with three separate thresholds. It would be interesting to address if the clusters found via the FT-IR fingerprints of human blood plasma reflect the "ageotypes" defined in [97]. Identifying clusters of individuals with similar biological characteristics, and therefore with similar spectra, could potentially boost the efficiency in the detection of diseases as grouping cases and controls accordingly would reduce the between-person variability which does not depend on the disease.

3.2.3 Comparison of FT-IR fingerprints of common parameters in KORA-FF4 and L4L

The FT-IR fingerprints of the common parameters analyzed for the KORA-FF4 cohort in section 3.2.2 are here compared with the ones of the healthiest individuals of the independent cohort L4L. The way the two cohorts have been sampled makes them, as well as the conclusions one can draw from these, very different. KORA is a population-based cross-sectional cohort and retains all natural correlations and distributions of conditions and common parameters of the general population. L4L, instead, is a case-control multi-clinical study and is therefore very specific to a target sub-population. In other words, the correlation coefficients found in L4L are not representative of a general population (Tables 3.6 and 3.7). For this reason, the analysis performed in the previous section, aiming at grasping the sources of the between-person spectral variability, would not have a general validity in the L4L cohort as in any other clinic-based study. This is very unique for the KORA study which makes it extremely valuable. On the other hand, comparing the IR fingerprints of the common parameters in independent cohorts allows the identification of parameter-specific IR features.

The L4L cohort has been collected in the frame of a cancer study, but in this dissertation, only the cancer-free patients are considered. The common medical conditions known for this cohort and in common with the KORA-FF4 study are four: diabetes, heart disease, hypertension and asthma (Table 3.4). Therefore, to guarantee a fair comparison, the cohorts of healthiest individuals are redefined as non-symptomatic (NSP*) only for these conditions for both KORA-FF4 and L4L. This returns a larger cohort of about 900 NSP* individuals for KORA-FF4 compared to the NSP/NGT previously investigated and allows the comparison of the IR fingerprints of the common parameters for a different definition of "healthy". Since the NSP* individuals are more numerous, they allowing a finer distinction for each parameter without extensively reducing the number of cases and controls. The thresholds defined in Table 3.2 of section 3.2.2 are therefore re-defined as reported in Table 3.5. Many values of inflammation and alcohol consumption are unknown in L4L (Figure 3.13b) and are not analyzed in this cohort.

L4L cohort					
Cohort	n. cases	M/F	Age	BMI (kg/m^2)	Smokers / not smokers
All	621	0.6	54.6 ± 14.6	24.7 ± 6.2	0.2
NSP*	439	0.4	51 ± 13.7	23.7 ± 5.9	0.2
Diabetes	30	1.7	67 ± 13.7	28.8 ± 5	0.3
Heart disease	46	1.9	71 ± 10.9	28 ± 6.3	0.1
Hypertension	131	1.1	64.9 ± 11.7	27.6 ± 6.3	0.2
Asthma	29	0.6	53.6 ± 13.3	27.4 ± 6.8	0.1

Table 3.4: Description of the L4L cohort. The table shows the number of individuals positive for each known common medical condition and of the healthiest non-symptomatic individuals (NSP*). For each cohort, the average values of age and BMI are reported together with the ratio between the number of male and female individuals and the one between active smokers and non-active smokers. Since L4L is not a cross-sectional population-based cohort, the distribution of show is different from the one seen for KORA-FF4 because it does not represent a general population. The NSP* individuals have the lowest average age and BMI compared to the other cohorts. Acronyms: NSP* - non-symptomatic individuals; BMI - body mass index; M/F - males-to-females ratio.

Parameter	Control cohort	Case cohort 1	Case cohort 2	Case cohort 2
Gender	females	males		
Age	<55 years old	> 54 years old		
BMI (kg/m^2)	< 25	25 - 29.9	> 30	
Smoking status	never smokers	ex-smokers	smokers	
CRP (mg/L)	< 2.5	2.5-5	5-7.5	> 7.5

Table 3.5: KORA-FF4 and L4L case/control cohorts for each common parameter among healthy individuals. The known parameters in common between KORA-FF4 and L4L are here analyzed in case-control studies. The thresholds used to identify these cohorts, which are based on different risk categories, are here listed for each parameter. The BMI is classified according to the WHO thresholds defining a BMI below 25 kg/m^2 as normal weight, between 25 and 29.9 kg/m^2 as preobese and above 30 kg/m^2 as obese. Acronyms: BMI - body mass index; WHO - World Health Organization; CRP - C-reactive protein.

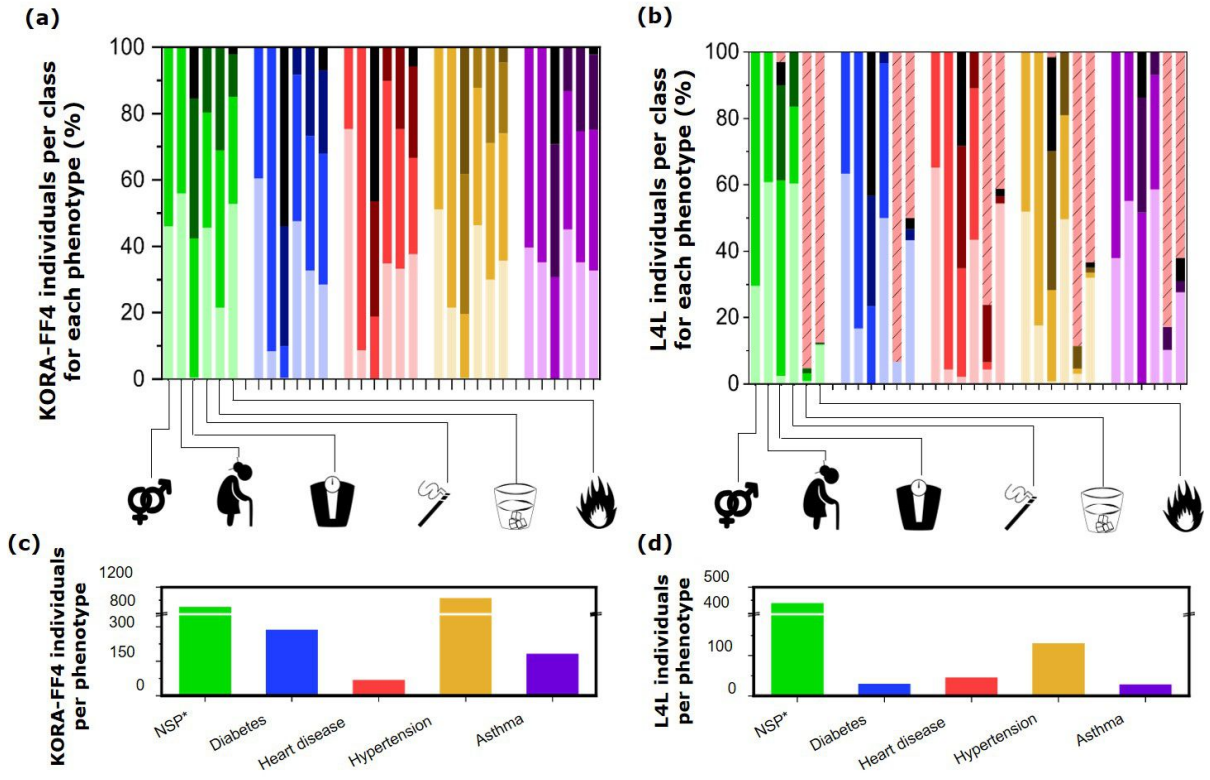

Figure 3.13: Description of the KORA-FF4 and L4L cohorts. (a) The percentages of individuals for each common parameter are reported for KORA and (b) L4L. The bars are colored based on the diseases (panels (c) and (d)) and the colors associated with the parameters in each bar go from light to dark according to the following order: gender (males, females); age (<55 years old, > 54 years old); BMI (underweight, normal weight, preobese, obese); Smoking status (non-, ex-, active smokers); alcohol intake (no alcohol consumption, < 10 g/day, above 10 g/day); inflammation (no, low, medium, high). The red barred columns represent the percentage of unknown values, showing that only a few values are known for alcohol consumption and inflammation for the L4L cohort. (c) The number individuals positive for each medical condition are reported for KORA (Table 3.1) and (d) L4L (Table 3.4). Acronyms: BMI - body mass index.

Besides for the individuals with high inflammation (18 cases), all the NSP* sub-cohorts include more than 130 people each. The percentages of individuals for each of these sub-cohorts are reported in Figure 3.13a and b for KORA-FF4 and L4L respectively. The common parameters are slightly differently distributed among the NSP* individuals of the two cohorts. For example, while in KORA-FF4 the number of males and females are the same, in L4L there is a small prevalence of males. Moreover, normal weight and preobese and obese individuals are about 50% each in KORA-FF4, while they are only 35% in L4L. A similar comparison is possible for all other conditions. The distributions (Figure 3.13a, b) and correlations (Tables 3.6 and 3.7) of the common parameters differ between the two cohorts highlighting that L4L and KORA-FF4 do not represent the same population.

χ^2 **p-values (KORA-FF4)**	Gender	Age	BMI
BMI	$7.5 \cdot 10^{-13}$	$2.8 \cdot 10^{-11}$	
Smoking status		$1.8 \cdot 10^{-10}$	$5.6 \cdot 10^{-4}$
Alcohol consumption	$2.2 \cdot 10^{-44}$		
CRP	$5.7 \cdot 10^{-3}$	$9.6 \cdot 10^{-6}$	$4.9 \cdot 10^{-16}$

Table 3.6: KORA-FF4 correlations between common parameters. The χ^2 derived p-values below 0.01 are reported for each parameter investigated. The correlations between smoking status and age and between smoking status and BMI are found only for KORA-FF4 but not for L4L cohort (Table 3.7). Acronyms: BMI - body mass index; CRP - C-reactive protein.

χ^2 **p-values (L4L)**	Gender	Age
Age	$2.1 \cdot 10^{-9}$	
BMI	$5.8 \cdot 10^{-8}$	$1.8 \cdot 10^{-4}$

Table 3.7: L4L correlations between common parameters. The χ^2 derived p-values below 0.01 are reported for each parameter investigated. The correlations between age and gender are found only for L4L but not for KORA-FF4 cohort (Table 3.6). Acronyms: BMI - body mass index.

The binary classification outcomes of the common parameters in KORA-FF4 are very similar for NSP* compared to the ones found for NSP/NGT cohort (Figure 3.10 and Figure 3.14a, b). It is noticeable that the definition of "healthy" does not introduce severe differences in the outcomes of binary classifications, most probably because of the strong impact of the common parameters on the IR fingerprints discussed in the previous section. The binary classification of all parameters in the KORA-FF4 cohort returns increasing AUCs for higher "risk" levels, e.g. for the case cohort 4 compared to the case cohort 1 (Table 3.5). The AUC of obese individuals is higher than when they are considered together with preobese ones, as in section 3.2.2. The same is true for inflammation, for which individuals with medium-to-high inflammation were previously classified together, with people with no or low inflammation. The binary classification of active smokers results in the same AUC when classified with never smokers alone or together with ex-smokers as in the previous section. This is because never and ex-smokers are indistinguishable via FT-IR spectroscopy (AUC = 50%).

The SVM coefficients found for NSP* individuals of KORA-FF4 and L4L are very similar (Figure 3.14b and d respectively). In general, the parameters giving higher AUCs have more features in common between the two cohorts. This highlights the strong stability of the differential fingerprints of common parameters independently of the nature of the analyzed cohort, which is not the case for the analysis of diseases (see chapter 4). Despite the high similarities of the SVM coefficients, the binary classification efficiencies are slightly different for KORA-FF4 and L4L. For example, active smokers are better classified in KORA-FF4 giving an AUC of 78% against the 60% of L4L. Moreover, the classification efficiencies of preobese and older individuals in L4L are higher than in KORA-FF4, going from 65% to 80% for age and from 66% to 72% for BMI. This could be due to the different chemical composition of the two biofluids and would therefore mean that the impact of smoking is stronger in human blood plasma compared to serum an vice versa for age and BMI.

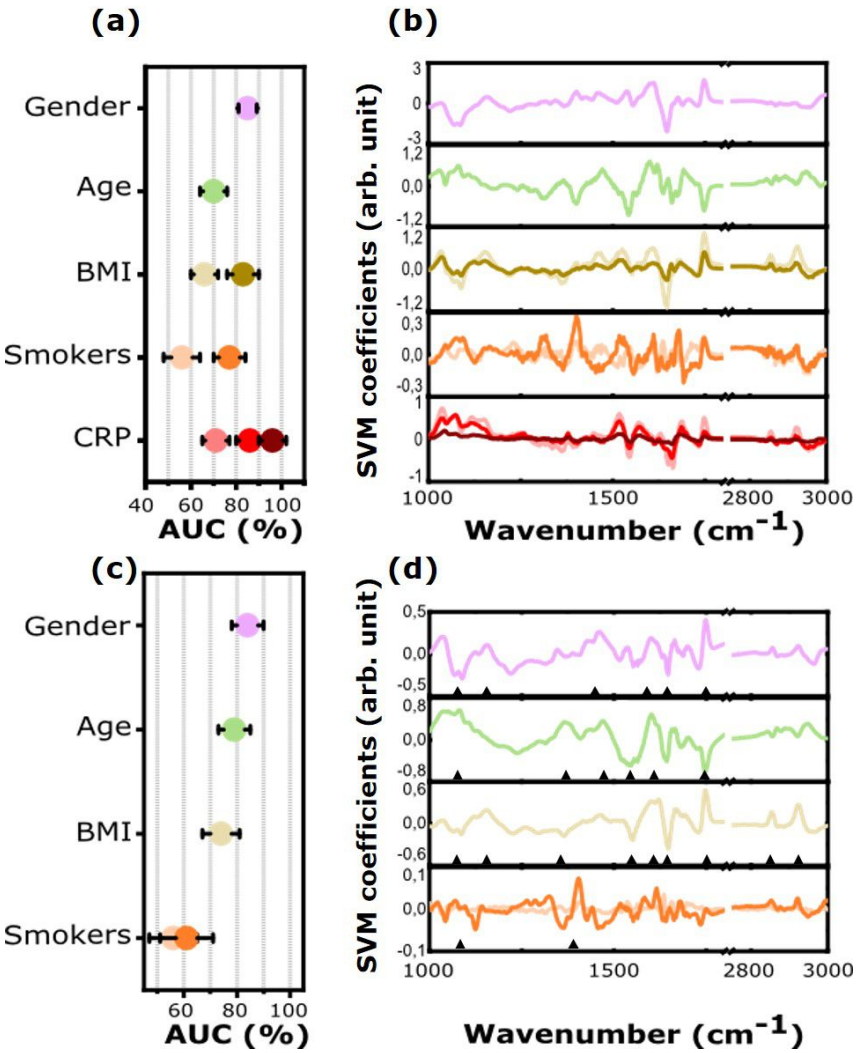

Figure 3.14: SVM binary classification of common parameters on the FT-IR spectra of healthy individuals in KORA-FF4 and L4L. (a) The AUC of the SVM binary classifications of NSP* individuals of KORA-FF4 return comparable results as for the previous definition of "healthy" (NSP/NGT, Figure 3.10) with increasing AUCs for higher risk levels (Table 3.5). (b) The SVM coefficients of the NSP* individuals of KORA-FF4 report the most important features in the corresponding binary classification. (c) The AUC of the NSP* individuals in L4L are slightly different than for KORA-FF4, with active smokers better classified in KORA-FF4, which would highlight a stronger effect on plasma compared to serum, and preobese and older individuals better classified in L4L. However, age correlate with gender in L4L but not in KORA-FF4 (Figure 3.13a), which could be the reason for the higher AUC observed in L4L compared to KORA-FF4. (d) The SVM coefficients of NSP* individuals of L4L are very comparable with the ones found for KORA-FF4. The black triangles highlight the features in common between the two cohorts. Acronyms: SVM - support vector machine; AUC - area under the curve; NSP* - non-symptomatic individuals; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; BMI - body mass index; CRP - C-reactive protein.

In conclusion, the different nature of the two cohorts leads to different mutual correlations between the common parameters influencing their classification efficiencies. However, most importantly, their signatures are highly comparable highlighting their robustness in independent different cohorts as well as in different biofluids (human blood plasma and serum). In the next chapter, it will be shown that this is not true for the fingerprints of diseases. This is another evidence of the strong impact of common parameters on the infrared spectra of human blood serum and plasma and, therefore, of the importance of their characterization before disease detection.

3.2.4 FRS and FT-IR fingerprints of common parameters in KORA-FF4

The FT-IR fingerprints of the common parameters in the KORA-FF4 cohort analyzed in section 3.2.2 are here compared to the ones found via FRS spectroscopy to validate the potential of this recently developed technique at its very first stages. To this end, since the spectral coverage of FRS is limited to 1000 - 1500 cm^{-1}, the FT-IR spectra are re-evaluated in the same spectral range for a proper comparison. The FRS data in the time domain are the EMFs introduced in section 3.1.2 and the corresponding data in the frequency domain are calculated as "absorption spectra" after applying the optimal HTPF (0.14 - 7 ps; Figure 3.15a, b), as described in section 3.1.3.

The FRS data have a strong dependence from the measurement day, compensated in the preprocessing via interference correction (section 3.1.2; Figure 3.7a, b). The PCA of samples and QCs together shows no signs of such dependence after applying the optimized preprocessing protocol (Figure 3.7b). However, if only the samples are considered, the dependence on the measurement day is attenuated but evident. This is because the interference correction preprocessing is based on the assumption that the refractive indexes of QCs and samples are the same. This is, of course, not fully true. The slightly different chemical compositions between samples and QCs, and even between samples, lead to different optical properties. For this reason, such correction cannot completely remove the day-to-day dependence, still visible in the FRS data both in the time domain and in the frequency domain. In particular, the day-to-day dependence affects PC1 in the frequency domain and PC2 in the time domain (Figure 3.15a, b), which is evident from the signature at 350 fs in LV2 (Figure 3.15e). Only one PC shows a dependence from the measurement day, highlighting that this effect is reduced by the optimal preprocessing. However, removing the affected PC returns lower AUCs for the binary classifications of both diseases and common parameters (data not shown). The removal of one of the first principal components is indeed not recommended as, besides unwanted sources of noise, they encode most of the valuable biological information. Aware of the need for improvements on the experimental side, this dataset still provides an important first validation of the potential of FRS fingerprinting.

The PCA of FT-IR with reduced spectral coverage and of FRS in both domains show that the first 5 PCs explain 90-95% of the total variance (Figure 3.15c), with PC1 addressing about 50% for both techniques in the frequency domain and about 45% for EMFs in the time domain. The cumulative explained variance is comparable between the two techniques. The loading vectors of FT-IR and FRS in the frequency domain are rather different, but most of the major features (in absolute value) are present in both approaches (Figure 3.15d).

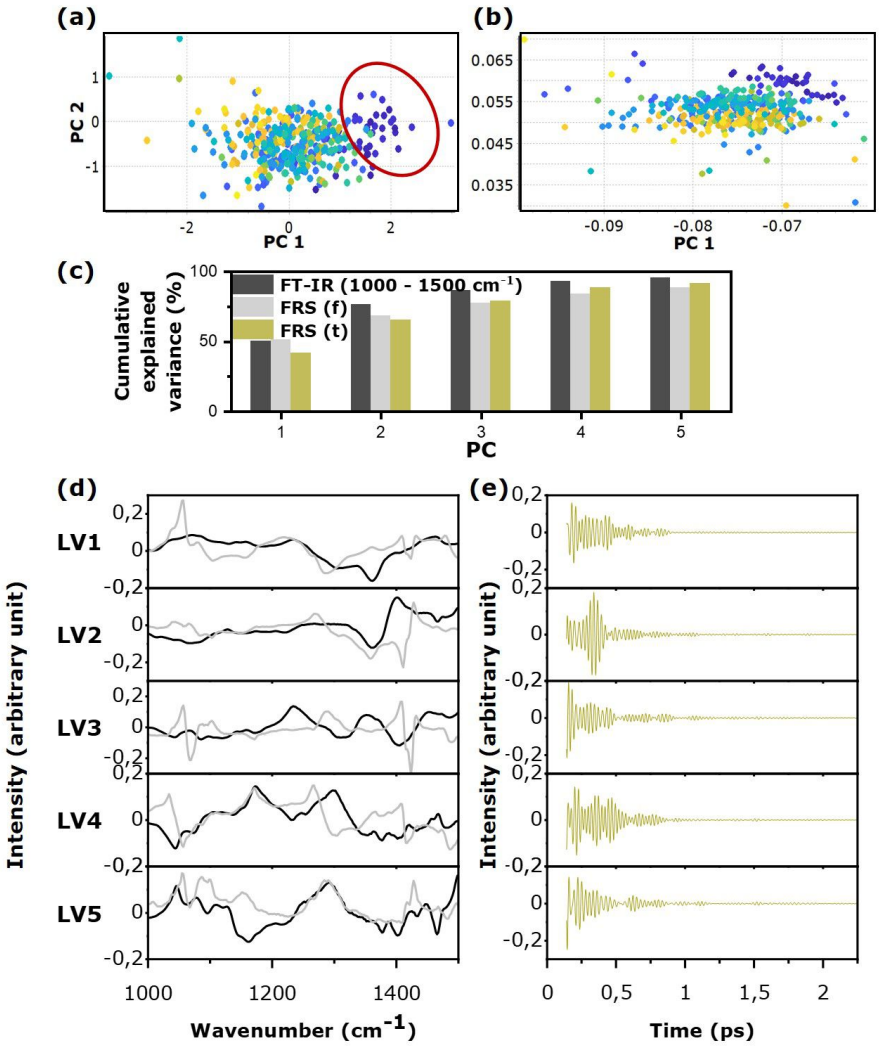

Figure 3.15: PCA analysis of the FT-IR reduced spectra and FRS data of healthy individuals in KORA-FF4. (a) The PC1/PC2 plots for samples measured via FRS in the frequency and (b) in the time domain (colored by day) show that the day-to-day dependence is not completely removed and affects PC1 and PC2 in the two cases respectively. (c) The cumulative explained variance of the first five PCs for the FT-IR reduced spectra is comparable with the one of the FRS data in the frequency and time domain. (d) The corresponding LVs of FT-IR reduced spectra and the FRS data in the frequency domain have similar shapes. (e) The LVs of FRS in the time domain show that the interference pattern around 350 fs, which causes the day-to-day dependence, affects only LV2 after applying the optimized preprocessing protocol. Acronyms: PCA - principal component analysis; FRS - field-resolves spectroscopy; PC - principal component; LV - loading vector.

The unsupervised analysis of FT-IR full-spectra has shown the role of lipids and proteins on the between-person spectral variability (section 3.2.2). A similar analysis with reduced spectral coverage provides a better understanding of the role of carbohydrates and protein glycosylation. The elbow method and the silhouette scores define the optimal number of clusters as two (Figure 3.16a-c) and identify the best-performing unsupervised method (Figure 3.16d-f). As in full-spectra, FT-IR returns similar silhouette scores for K-means and GMM (covariance type full), the last giving the highest silhouette score also for FRS (with spherical covariance in the

time domain). The clusters identified in the PC1/PC2 plots cannot be visually compared (Figure 3.16g-i). The individuals in the two clusters are completely different in the three cases because the day-to-day dependence influence the clustering of FRS data, especially in the frequency domain (red circle in Figure 3.15a and Figure 3.16h refer to the same datapoints; note: PC2 is inverted in the two figures).

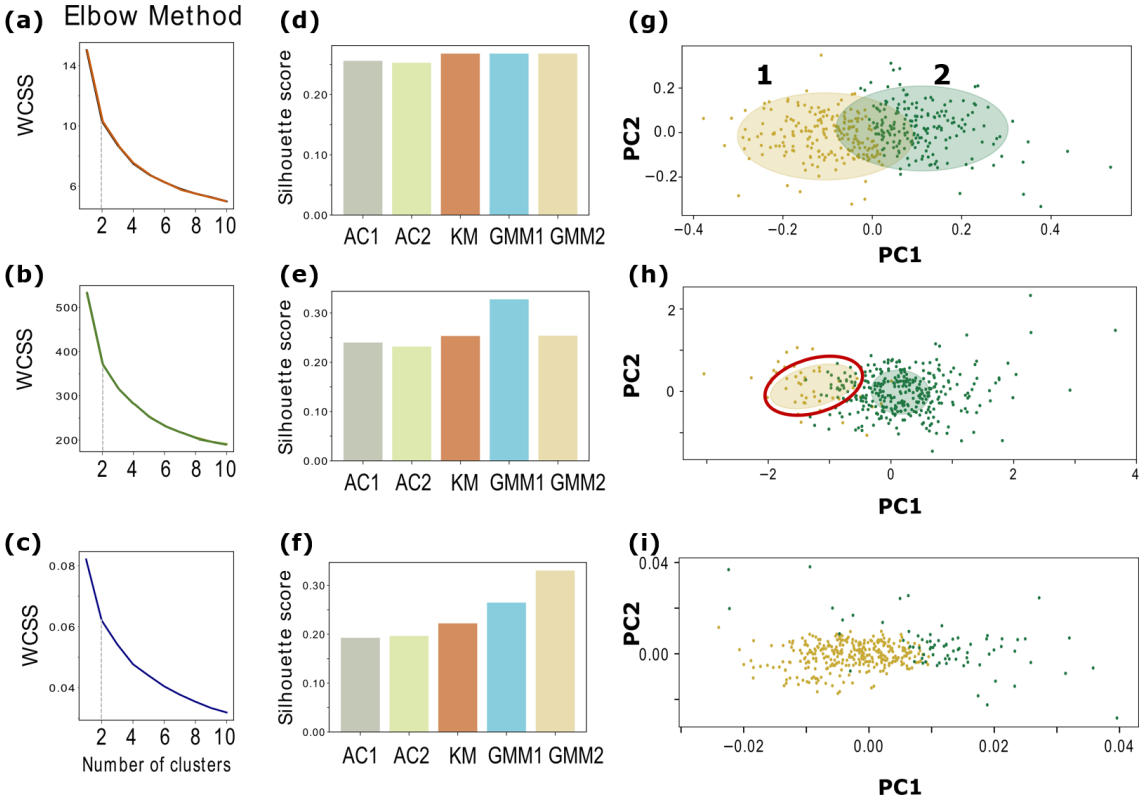

Figure 3.16: Unsupervised clustering analysis of the FT-IR reduced spectra and FRS data of healthy individuals in KORA-FF4. (a) The elbow method shows the WCSS against the number of clusters for FT-IR in the reduced spectral coverage (1000 - 1500 cm^{-1}), (b) FRS in the frequency domain and (c) FRS in the time domain showing that two is the optimal number of clusters. (d) The silhouette scores are reported for the unsupervised algorithms applied on the PCs explaining 99% of the total variability calculated for the NSP/NGT cohort for FT-IR with reduced spectral coverage (1000 - 1500 cm^{-1}), (e) FRS in the frequency domain and (f) FRS in the time domain. The graph shows that K-means and GMM have the best performances for FT-IR, while GMM1 and GMM2 have the best performance for the FRS data in the frequency and time domain respectively. (g) The PC1/PC2 plot colored according to the clusters identified via the respective best performing algorithm (panels (a)-(c)) for FT-IR in the reduced spectral coverage, (h) FRS in the frequency domain and (f) FRS in the time domain. The clusters identify completely different datapoints in the three cases. In particular, the clusters found for the FRS data are strongly affected by the day-to-day dependence in both the time and frequency domain as it shows the comparison with Figure 3.15a (the red circle highlights the same datapoints in the two figures; PC2 is inverted). Acronyms: FRS - field-resolved spectroscopy; WCSS - within-cluster sum of squares; PC - principal component; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; AC - Agglomerative clustering (1 - affinity = 'euclidean', linkage = 'ward'; 2 - affinity = 'cosine', linkage = 'average'); KM - K-means; GMM - Gaussian Mixture Model (1 - covariance type = 'full'; 2 - covariance type = 'spherical').

The distributions of the common parameters in the two clusters found for FT-IR truncated spectra highlight that there is a small unbalance in the distribution of males and females between the two clusters and that most of the people with high inflammation are in cluster 1 (Figure 3.17a). The average CRP concentration in cluster 1 is 2 ± 3 mg/L, reaching up to 25 mg/L, while average and maximum CRP concentrations are smaller in cluster 2, respectively 1 ± 1 mg/L and 8mg/L. In this spectral region, the main signatures come from carbohydrates and protein glycosylation; the source of the between-person variability of these biomolecules is affected by a different set of common parameters compared to the ones observed in full-spectrum, in particular by gender and inflammation (section 3.2.2). These outcomes agree with the correlations observed in the analysis of full-spectra between gender with PC3 and, together with inflammation, with PC4 and PC5 (Figure 3.12c). Indeed, the LV3, 4 and 5 found for FT-IR in full-spectra are equivalent to the first three LVs found for the reduced spectra (Figure 3.17b).

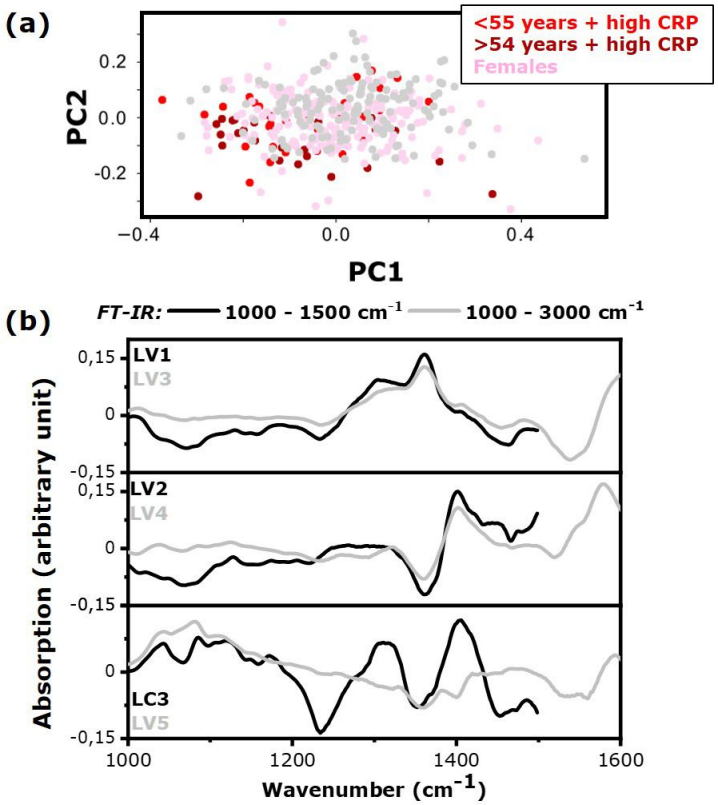

Figure 3.17: PCA analysis of the FT-IR full and reduced spectra of healthy individuals in KORA-FF4. (a) The PC1/PC2 plot of NSP/NGT individuals colored by specific sub-cohorts highlights the importance of gender and inflammation in the clustering if the spectral region is reduced to 1000 - 1500 cm^{-1} (Figure 3.16g). (b) The first three LVs for FT-IR with reduced spectral coverage are comparable with the LV3, 4 and 5 in full-spectra. This explains why gender and inflammation, which are shown to correlate with PC3, PC4 and PC5 in full-spectra (Figure 3.12c), are the most relevant parameters in the clustering of the FT-IR reduced spectra. Acronyms: PCA - principal component analysis; PC - principal component; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; LV - loading vector.

Cluster	Males	Females	Low CRP concentration	High CRP concentration
1	43.8%	56.9%	49.7%	**85.0%**
2	56.2%	43.5%	50.3%	**15.0%**

Table 3.8: GMM clusters found for the FT-IR reduced spectra of healthy individuals in KORA-FF4. The distributions of the common parameters that correlate with PC3, PC4 and PC5 in FT-IR full-spectra are the only unbalanced ones between the two clusters. In particular, there is a small unbalance in the percentage of all males and all females between the two clusters. The individuals with low inflammation are equally distributed between the two clusters, while most of the people with high inflammation are in cluster 1 (Figure 3.17a). Acronyms: GMM - Gaussian Mixture Model; PC - principal component; CRP - C-reactive protein.

The SVM binary classifications of common factors return similar AUCs with both techniques (Figure 3.18a). The classification of both active smokers and medium-to-high inflammation levels are based on a few cases (85 and 55 individuals respectively) leading to noisy SVM coefficients. In general, the SVM coefficients of FRS in the frequency domain are noisier than the ones found via FT-IR and the comparison is not straightforward (Figure 3.18c).

The SVM classification of EMFs in the time domain is performed by sliding different temporal windows along the whole time trace and evaluating the AUC for each of them to identify the temporal range delivering the highest efficiency (see section 2.3.2). Figure 3.18b shows the AUC along with the whole time window for the binary classification of inflammation. The yellow area identifies the temporal window delivering the highest AUC, also depicted in Figure 3.18d which shows the SVM coefficients of each parameter plotted in the corresponding time window delivering the highest classification efficiency. As it can be seen from panel b, the optimal time window is only the one around the maximum AUC, but the classification efficiencies might be high also for larger windows and at different times. Therefore, the SVM coefficients reported in panel d shows only the most important features in the binary classification, despite those are not the only one contributing. Different temporal windows carry the information relevant to each parameter. In other words, each parameter is associated with a unique temporal fingerprint. At longer times the signal gets smaller, the AUCs tend to drop and the overfitting rates rise (Figure 3.18b). The optimal preprocessing described in section 3.1.2 is the only one able to consistently reduce overfitting in SVM binary classifications of EMFs, thus making the information in the whole time window accessible. However, if standard scaling is applied before PCA, overfitting extensively affects the classifications in the whole time window.

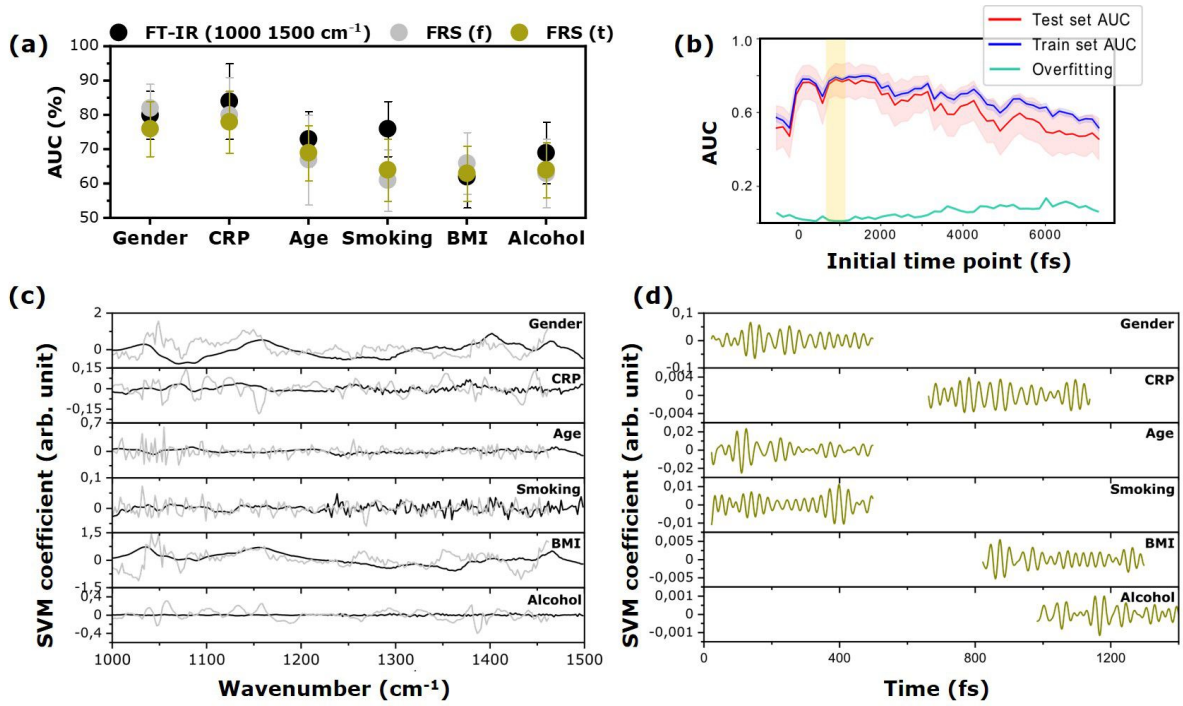

Figure 3.18: SVM binary classification of common parameters on the FT-IR reduced spectra and the FRS data of healthy individuals in KORA-FF4. (a) The AUCs are similar for the two techniques in both the time and frequency domain. (b) The AUC derived for the EMFs along with the whole time window for the binary classification of CRP highlights that despite the optimal time window (yellow area) identifies the best time window, the AUC can be higher than 50% also at other times. (c) The SVM coefficients corresponding to panel (a) for FT-IR reduced spectra and FRS in the frequency domain are not easily comparable because the FRS data return noisy coefficients. (d) The SVM coefficients derived from the EMFs data are shown in the optimal time windows and highlight the features contributing the most to the corresponding binary classification. The SVM coefficients are very different for each parameter and show that FRS provides more parameter-specific signatures than for FT-IR for which all the features overlap. Acronyms: SVM - support vector machine; FRS - field-resolved spectroscopy; AUC - area under the curve; EMF - electric-field-resolved molecular fingerprint; CRP - C-reactive protein; BMI - body mass index; FRS(f) - FRS signal in the frequency domain; FRS(t) - FRS signal in the time domain.

Overall, considering the day-to-day dependence affecting the FRS data, reaching comparable classification efficiencies as a state-of-the-art FT-IR spectrometry is a promising starting point. The signatures of the common parameters in the EMFs are highly distinguishable highlighting how FRS has the potential to deliver highly parameter-specific signatures compared to FT-IR spectroscopy, or in the frequency domain in general, for which the features of other correlated factors might overlap. However, further developments on the technical and analytical sides are still needed and are currently being implemented.

3.3 Concluding remarks

In this chapter, principal component analysis (PCA), supervised and unsupervised machine learning algorithms are used to address the impact of the common parameters on the FT-IR and FRS data of the healthiest individuals of KORA-FF4 and L4L cohorts, with a particular interest in their effect on the between-person spectral variability.

As a first step, the technical noise affecting the measurements of biological samples is investigated using hundreds of replica of the same human serum sample measured together with the KORA-FF4 samples. An optimized preprocessing protocol has been identified for both FT-IR and FRS. The measurement campaign of KORA-FF4 is indeed among the largest published studies based on FT-IR and the first large sample set measured via FRS spectroscopy. The characterization of the technical noise highlights that FRS data are affected by a strong day-to-day dependence due to the variable ambient pressure acting on the top of the measurement cell (cuvette) influencing its thickness. This effect is reduced with the optimized preprocessing, but the dependence on the measurement day still affects the first principal components of the FRS data both in the time and frequency domain introducing an important bias in the unsupervised clustering analysis.

Despite the residual day-to-day dependence, FRS spectroscopy returns comparable classification efficiencies as a state-of-the-art FT-IR spectrometer. FRS has been already shown to reach lower LOD compared to FT-IR and to be suited for measuring optically thick samples for which FT-IR is cannot be applied [52]. Moreover, compared to any technique based on the frequency domain, the resolution in time allows disentangling contributions that would overlap in the absorption spectra, which gives FRS the important advantage of identifying very different signatures for the common parameters investigated (Figure 3.18d). Further technical developments are currently being implemented in the FRS system to make it more robust to the daily changes as well as to boost its classification efficiency by expanding the spectral coverage and lowering the LOD.

The analysis of FT-IR spectra addresses the origin of the known between-person spectral variability in a large cross-sectional population-based cohort for the first time to the best of our knowledge. In FT-IR full-spectra, age and inflammation are the main sources of the between-person spectral variability. In particular, age influences the signature of lipids (1750 - 3000 cm^{-1}, the highest source of variability), in agreement with what reported in the literature [93, 94, 99, 100], and proteins (1250 - 1750 cm^{-1}, the second-highest source of variability), while inflammation affects only the second. Aging is connected with inflammation [97], which is why most of the age-related diseases share an inflammatory pathogenesis [98]. The connection between age and inflammation might explain the impact of age also on the AGR signature typically attributed to inflammation [61, 92]. Using unsupervised clustering to define age and CRP concentration thresholds does not provide the same clusters as the ones identified via unsupervised methods. This could be due to the influence of other unknown parameters on the IR fingerprints of healthy individuals or to the fact that different people go through aging at different rates and, potentially, via different mechanisms [97]. Therefore, identifying individuals with similar blood composition based on their age might be inefficient and a more proper way could be based on clustering them according to their "effective age" [97]. To address this issue, it is important to further analyze the clusters of FT-IR fingerprints and look for connections between these clusters and the "ageotypes" defined from other studies [97] as

well as to investigate the IR fingerprint of other parameters not considered in this analysis (e.g. hormonal status, diet, medications, cholesterol levels, ...). The importance of identifying clusters of individuals with similar biological characteristics, and therefore with similar spectra, is due to the potential advantage in grouping cases and controls accordingly to reduce the between-person variability of both cohorts boosting the efficiency of IR fingerprinting in the detection of diseases.

The unsupervised clustering of the same FT-IR spectra in the spectral coverage of FRS (1000 - 1500 cm^{-1}) shows that the main parameters connected with the variability in this spectral region are inflammation and gender. In particular, the main signatures at these frequencies come from carbohydrates and protein glycosylations, showing how gender and inflammation influence the variability of these biomolecules. The connection between CRP and protein glycosylation can have multiple origins. Inflammation and protein glycosylation are indeed reported to influence each other: inflammation affects the glycosylation and, therefore, the functional capacity of antibodies [101], while changes in protein glycosylation can modulate the inflammatory responses of the body [102]. Gender, age, BMI and smoking status have been shown to affect protein glycosylation too [103–105]. Males and females have been shown to have different glucose tolerance, probably due to the sexual dimorphism in the hepatic insulin action [106], which could explain the association of gender with the spectral signature of carbohydrates. However, it is important to stress that this part of the spectra accounts for only about 2% of the total variability in full-spectra. This is in agreement with what was reported in the literature, according to which age-associated changes in the human blood plasma composition are more pronounced than the ones related to gender and other factors, in particular for what regards changes in protein, lipid metabolism and oxidative stress [107].

The comparison of the binary classification of common parameters in KORA-FF4 and an independent cohort highlights that the different correlations between the common parameters lead to different classification efficiencies. Chapter 4 will address the importance of matching case and control cohorts to reduce disease-unspecific contributions due to correlations of the disease under investigation with common parameters or other medical conditions.

Despite the different AUCs, the signatures of each of the common parameters analyzed are highly comparable in the two cohorts, which does not hold for the signatures of medical conditions (see chapter 4). This proves that the influence of common parameters on the IR fingerprints of human blood plasma is even stronger than the spectral signature of diseases. Moreover, both the cumulative explained variance and loading vectors are almost identical for healthy individuals and the whole population in KORA-FF4 (Figure 3.9a-b), showing that the molecular origin of the between-person spectral variability is the same in the two cases. The higher between-person spectral variability found for endpoint and intermediate medical conditions (Figure 3.9b) can be explained as the simple consequence of the higher average age and inflammatory levels of symptomatic individuals (Table 3.1). These observations stress the importance of providing a consistent characterization of the infrared spectral signatures of common parameters in a large cross-sectional population-based cohort before applying any IR spectroscopy for the detection of medical conditions.

Chapter 4

FT-IR and FRS fingerprinting for disease diagnosis

Human blood serum and plasma provide real-time information on the human phenotypes and health status in a minimally invasive fashion [108]. Vibrational spectroscopy records a snapshot of all molecular constituents simultaneously providing IR fingerprints that correlate with any change induced by phenotypes or medical conditions on both concentration and structure of different biomolecules [14, 15, 70, 71, 109]. Recent examples of diagnosis via Raman and FT-IR spectroscopy on human blood biofluids have been reported for several diseases [110–120]. However, these analyses are usually performed on clinic-based case-control cohorts. In this chapter, FT-IR and FRS spectroscopy are applied on the 2500 individuals of the cross-sectional population-based KORA-FF4 for the spectral detection of common medical conditions. In particular, the focus is on the following medical endpoints: diabetes, hypertension, heart attack, asthma, high blood lipids, chronic obstructive pulmonary disease (COPD), as well as on individuals who had cancer or experienced episodes of stroke and on the intermediate condition of prediabetes. The cross-sectional population-based KORA-FF4 cohort provides the unique advantage to allow identifying the fingerprints of each condition in the general population, with the influence of all naturally correlated common parameters and comorbidities. The importance of matching cases and controls for common parameters and comorbidities to assess the robustness of the approach and to isolate the disease-specific features are additionally evaluated for the FT-IR fingerprints of diabetes, hypertension, heart disease and asthma via the independent cohort L4L. Moreover, a comparison of the diagnostic power of IR fingerprinting of each condition is addressed considering the influence of the statistical power, namely of the number of cases, of each classification. Before going into disease diagnosis, unsupervised clustering methods are applied on the whole KORA-FF4 population as performed in chapter 3 for the healthy (NSP/NGT) individuals to asses the impact of common parameters on the IR fingerprints of symptomatic individuals and identify the effect of each medical condition in the clustering. In section 3.2.4 it has been shown how the FRS data are influenced by the measurement day too strongly to obtain reliable results with unsupervised clustering (section 3.2.4). Therefore, this analysis is performed only on FT-IR in full-spectra.

4.1 Clustering of KORA-FF4 FT-IR fingerprints

Unsupervised clustering algorithms are used to identify groups of individuals with similar FT-IR spectra and, therefore, similar blood plasma biochemical composition. These methods allow addressing what are the most relevant common parameters and medical conditions affecting the between-person spectral variability. Applied to the healthiest sub-cohort of the population, unsupervised clustering has allowed identifying age and inflammation as the main factors affecting the between-person spectral variability in the full-spectral coverage (3.2.2). The same analysis is here applied to the whole population to address the effect of these parameters in the general population, including individuals symptomatic for the known endpoint and intermediate medical conditions.

As in section 3.2.2, the elbow method address three as the optimal number of clusters (Figure 4.1a). Silhouette scores are used to identify the best performing algorithm. While K-means and GMM had comparable performances for the healthy individuals, on the whole population K-means returns a higher silhouette score and is therefore selected as the optimal clustering method (Figure 4.1b).

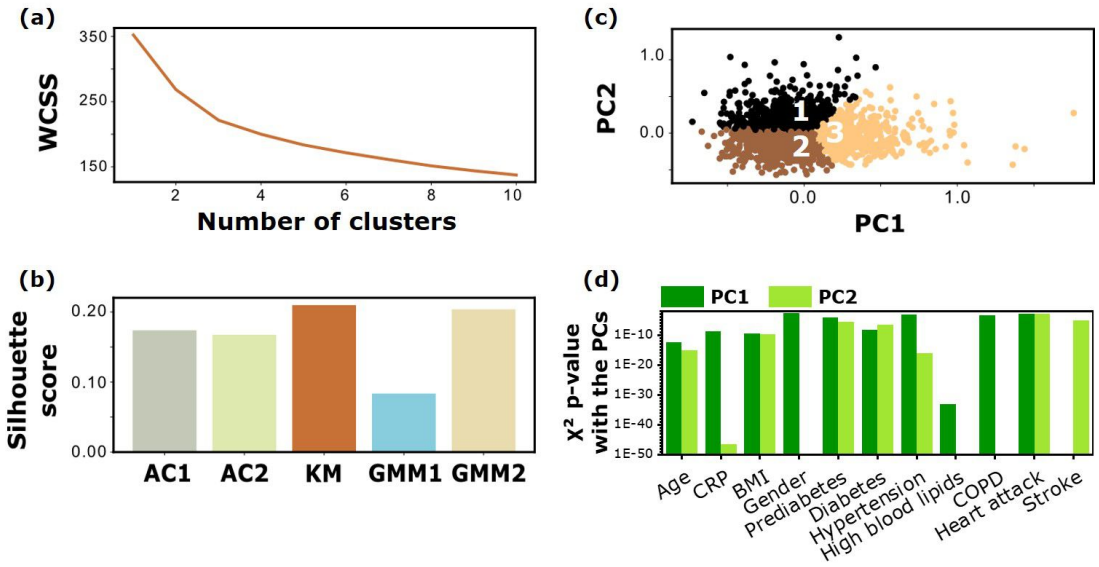

Figure 4.1: Unsupervised clustering analysis of the FT-IR spectra of the whole KORA-FF4 cohort. (a) The elbow method plots the WCSS against the number of clusters showing that three is the optimal number of clusters for this cohort. (b) The silhouette scores are calculated for the unsupervised algorithms applied and show that K-means is the optimal algorithm for this cohort. (c) The PC1/PC2 plot (colored according to the clusters found via K-means) shows that the three clusters separate along PC1 and PC2 which, therefore, address the main between-person spectral variability. (d) The χ^2 derived p-values show the correlations of common parameters and medical conditions with the first two PCs. Age and inflammation have a strong impact on PC1 and PC2 also for the whole population. High blood lipids ad hypertension have a strong correlation with PC1 and PC2 respectively. Acronyms: WCSS - within-cluster sum of squares; PC - principal component; CRP - C-reacve protein; BMI - body mass index; COPD - chronic obstructive pulmonary disease; AC - Agglomerative clustering (1 - affinity = 'euclidean', linkage = 'ward'; 2 - affinity = 'cosine', linkage = 'average'); KM - K-means; GMM - Gaussian Mixture Model (1 - covariance type = 'full'; 2 - covariance type = 'spherical').

The three clusters identified via K-means separate along PC1 and PC2 (Figure 4.1c), similarly as for the healthy individuals (Figure 3.11). It is important to remember that the LVs are identical for the healthy as well as for the whole cohort (Figure 3.12b), therefore the analysis brings to similar conclusions as in section 3.2.2. In particular, the distributions of the common parameters and medical conditions in the three clusters identified on the whole cohort are reported in Table 4.1. About 64% of the NSP/NGT individuals end up in cluster 2 which counts the youngest individuals and the ones with the lowest inflammation levels. The impact of age and inflammation on PC1 and PC2 are very strong also on the whole population (Figure 4.2c). Individuals with COPD as well as the ones who had an episode of heart attack or stroke and former cancer patients end up in cluster 1. Most of the people with high blood lipids are in cluster 3, while individuals affected by diabetes or prediabetes are almost equally distributed between clusters 1 and 3 (Table 4.1).

Parameter	Cluster 1	Cluster 2	Cluster 3
NSP/NGT	25%	64%	11%
M/F	1.1%	1.0%	0.8%
Age	63.9 ± 12.9	56.0 ± 11.7	61.9 ± 10.7
CRP (mg/L)	3.8 ± 6.3	1.3 ± 2.9	2.5 ± 3.2
Smokers	30%	40%	30%
Alcohol consumption (g/day)	14.1 ± 20.5	13.8 ± 17.5	16.9 ± 22.8
BMI (kg/m^2)	28.9 ± 5.4	26.3 ± 4.0	28.7 ± 5.0
Prediabetes	40%	22%	38%
Diabetes	37%	24%	39%
Hypertension	38%	32%	32%
High lipids	36%	39%	43%
COPD	39%	26%	35%
Asthma	34%	33%	34%
Heart disease	54%	33%	13%
Stroke	65%	15%	20%
Ex-cancer	44%	31%	24%
n. comorbidities	1.8 ± 1.4	1.2 ± 1.2	1.8 ± 1.2

Table 4.1: K-means clusters found for the FT-IR fingerprints of the whole KORA-FF4 cohort. The average values of common parameter as well as the ratio between the number of male and female individuals are reported for each cluster. Age and inflammation are the most unequally distributed parameters among the three cluster. In particular, the individuals in cluster 2 have the lowest average age and CRP concentration compared to the individual of the other clusters. The percentage of the individuals symptomatic to each known medical condition are reported for each cluster. In particular, the cases positive to the conditions correlating with age mostly end up in cluster 1, while most of the individuals with high blood lipids are in cluster 3. In general, the distributions follow the correlation with the first two PC2 (Figure 4.1d). Acronyms: NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; M/F - males-to-females ratio; CRP - C-reactive protein; BMI - body mass index; COPD - chronic obstructive pulmonary disease.

In summary, if cluster 2 is taken as a reference, the individuals in clusters 1 and 3 differ because of the higher average age and inflammation levels, which is expected because of the high similarity between the LVs of healthy individuals and the whole population (Figure 3.12b). In general, the conclusions derived from unsupervised clustering are the same that can be derived from PCA analysis: the first PCs address the main sources of variability, in particular, due to the spectral signatures of lipids and proteins which are affected by the common parameters and the medical conditions that correlate with these components (Figure 4.1d). For example, the higher percentages of individuals affected by hypertension in cluster 1 and of the ones with high blood lipids in cluster 2 are due to the correlations of these conditions with PC2 and PC1 respectively, which in other words means that they impact the features of the corresponding LVs.

The unsupervised clustering analysis reported addresses specifically the features and the effect of each common parameter connected with the between-person spectral variability, but it does not show how strongly each factor affects the IR fingerprints in general. To this end, SVM binary classifications are performed in the next section.

4.2 IR spectral detection of clinical endpoints

4.2.1 FT-IR fingerprints of medical conditions in independent cohorts

An increasing amount of research emphasizes the importance of matching cases and controls to remove potential bias from non-target covariates and to increase the specificity of the analysis [15, 82, 121–124]. The present study confirms the importance of matching case and control cohorts via analyzing the same medical conditions in two large independent cohorts. In particular, the FT-IR fingerprints of individuals affected by type II diabetes, hypertension, heart disease and asthma are compared for KORA-FF4 and L4L cohorts via SVM binary classifications with and without matching cases and controls for the common parameters and comorbidities correlated with each medical condition.

As introduced in section 3.2.3, the information and nature of the medical conditions compared in the two cohorts are slightly different. In particular, all parameters are self-reported from the participant to each study, except for type II diabetes for the KORA-FF4 cohort which has been addressed via OGTT. Moreover, the cohorts of individuals labeled as "heart disease" refer to people who had an episode of a heart attack in the KORA-FF4 cohort, while it is self-reported and generic for the L4L cohort. The number of cases symptomatic to each condition is also very different in the two cohorts (Figure 4.2a, b). In general, L4L is a smaller study compared to KORA-FF4 and the number of symptomatic individuals is less than 150. Moreover, KORA-FF4 is a cross-sectional population-based cohort and represents a very different part of the population compared to the case-control clinic-based L4L study. As a consequence, the correlation coefficients between common parameters and medical conditions for the two cohorts are different and only a few of the correlations are found in both cohorts (Tables 4.2 and 4.3). For example, hypertension correlates with gender and heart disease correlates with BMI only for the L4L cohort but not for the KORA-FF4 cohort. Since the two populations are completely independent and have a different nature, any signature found for the same medical condition in both studies is expected to be specific to that condition.

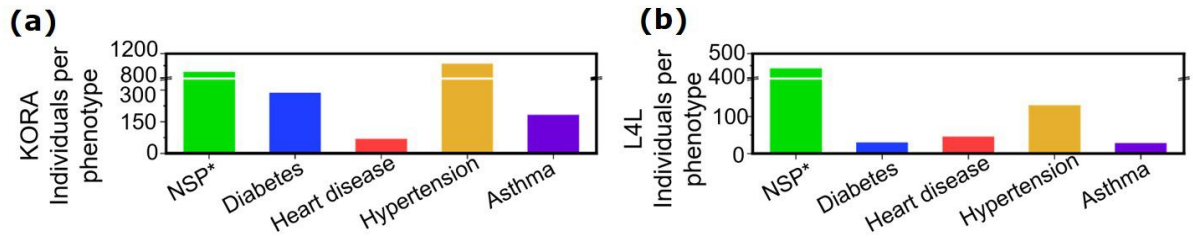

Figure 4.2: Medical conditions in the KORA-FF4 and L4L cohorts. (a) The number of individuals is reported for each medical condition in KORA-FF4 (Table 3.1) and (b) L4L cohorts (Table 3.4). Acronyms: NSP* - non-symptomatic (for diabetes, hypertension, heart disease and asthma).

KORA-FF4 cohort									
p-values	Predia-betes	Dia-betes	Hyper-tension	High blood lipids	Heart at-tack	Stroke	Former cancer	COPD	Asthma
Gender	$3.9 \cdot 10^{-4}$	$3.2 \cdot 10^{-5}$			$1.2 \cdot 10^{-5}$				
Age	$1.1 \cdot 10^{-15}$	$8.2 \cdot 10^{-28}$	$1.3 \cdot 10^{-49}$	$1.4 \cdot 10^{-15}$	$9.9 \cdot 10^{-7}$	$4.3 \cdot 10^{-7}$	$4.3 \cdot 10^{-7}$	$6.8 \cdot 10^{-6}$	
BMI	$1.7 \cdot 10^{-13}$	$1.1 \cdot 10^{-15}$	$1.0 \cdot 10^{-25}$	$3.5 \cdot 10^{-7}$				$8.9 \cdot 10^{-3}$	
CRP	$6.3 \cdot 10^{-6}$	$4.6 \cdot 10^{-6}$	$4.3 \cdot 10^{-7}$		$6.0 \cdot 10^{-3}$			$9.3 \cdot 10^{-4}$	
Diabetes	$8.2 \cdot 10^{-17}$								
Hyperten.	$3.1 \cdot 10^{-7}$	$7.1 \cdot 10^{-25}$							
High blood lipids		$2.3 \cdot 10^{-8}$	$2.6 \cdot 10^{-12}$						
Heart at-tack		$2.3 \cdot 10^{-11}$	$9.1 \cdot 10^{-7}$	$2.1 \cdot 10^{-7}$					
Stroke	$3.5 \cdot 10^{-3}$	$2.9 \cdot 10^{-4}$	$3.9 \cdot 10^{-6}$	$3.3 \cdot 10^{-4}$	$4.6 \cdot 10^{-3}$				
Former cancer		$1.5 \cdot 10^{-4}$	$9.2 \cdot 10^{-4}$						
COPD		$1.1 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$					$2.7 \cdot 10^{-47}$

Table 4.2: KORA-FF4 correlations between common parameters and medical conditions. The χ^2 derived p-values below 0.01 are reported for each parameter investigated. Acronyms: BMI - body mass index; CRP - C-reactive protein; COPD - chronic obstructive pulmonary disease.

L4L cohort			
p-values	Diabetes	Hypertension	Heart disease
Gender	$2.6 \cdot 10^{-3}$	$2.7 \cdot 10^{-5}$	$3.4 \cdot 10^{-5}$
Age	$3.4 \cdot 10^{-4}$	$1.7 \cdot 10^{-16}$	$3.6 \cdot 10^{-10}$
BMI	$1.5 \cdot 10^{-3}$	$6.6 \cdot 10^{-10}$	$5.6 \cdot 10^{-3}$
Hypertension	$3.1 \cdot 10^{-6}$		
Heart disease	$2.2 \cdot 10^{-3}$	$2.8 \cdot 10^{-10}$	

Table 4.3: L4L correlations between common parameters and medical conditions. The χ^2 derived p-values below 0.01 are reported for each parameter investigated. The correlations are slightly different to what observed for KORA-FF4. For example, hypertension correlates with gender and heart disease correlates with BMI only for the L4L cohort but not for the KORA-FF4 cohort. Asthma and smoking status do not significantly correlate with any other parameters or diseases. Acronyms: BMI - body mass index.

In a first analysis, the individuals non-symptomatic to the four medical conditions (NSP*, analyzed in section 3.2.3) are selected as controls for the SVM binary classification of each disease and cases and controls are not matched. The distribution of the common parameters for the individuals symptomatic to the four medical conditions in the KORA-FF4 and L4L cohorts can be found in Table 3.1 and 3.4 respectively. The classification efficiencies are similar for the same medical conditions in both cohorts and are above 50%. The highest AUC is found for diabetes, followed by heart disease, hypertension and asthma (Figure 4.3a, b). The AUCs found in the L4L cohort have smaller mean values and higher standard deviations as compared to the ones found for the KORA-FF4 cohort because of the smaller number of cases in the L4L study (see section 4.2.3). The SVM coefficients found for each medical condition are similar in the same cohort. For example, the coefficients found for diabetes, hypertension and heart disease classes of the KORA-FF4 cohort have a similar signature at low frequency (1000 - 1250 cm^{-1}, Figure 4.3c). Similarly, the coefficients found for the L4L cohort are highly comparable for hypertension and heart disease in the whole spectral range (Figure 4.3d). The similarities found between the SVM coefficients are due to the correlations between the four medical conditions. However, as seen before, different cohorts have different correlations because they do not represent the same part of the population (Tables 4.2 and 4.3). If the case and control cohorts are not matched, non-target medical conditions and common parameters correlating with the target disease are more numerous among the cases compared to the controls, thus influencing the binary classification outcomes. As a result, correlating medical conditions share common features in the SVM coefficients obtained via binary classifications with unmatched controls (Figure 4.3c, d).

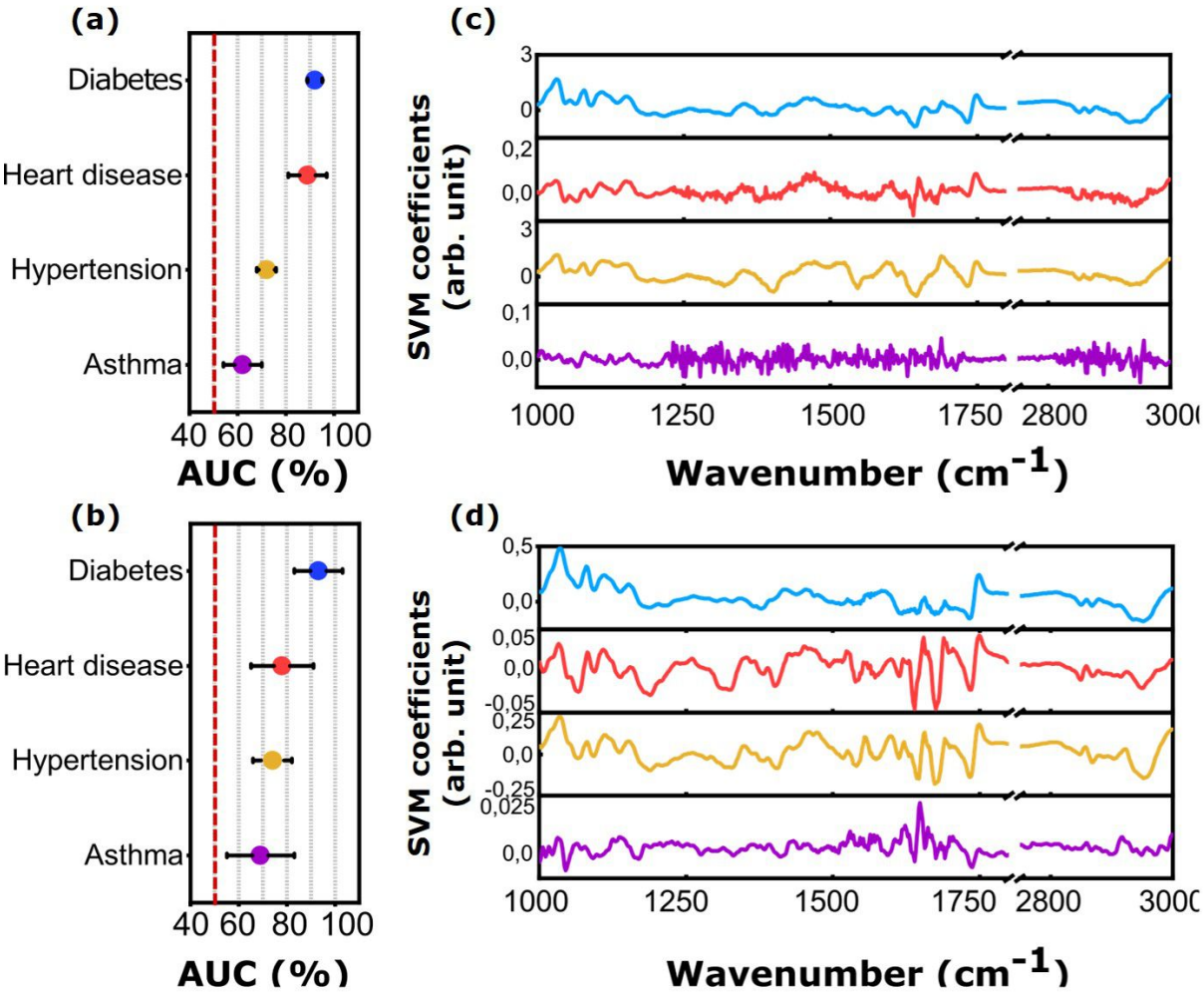

Figure 4.3: SVM binary classification of four common medical conditions with NSP* unmatched individuals on the FT-IR spectra of KORA-FF4 and L4L cohorts. (a) The AUC scores found for KORA-FF4 and (b) L4L cohorts have the same trend for the four medical conditions and are always above 50%. The number of cases is smaller for all conditions in the L4L cohort for which the AUCs have higher standard deviations compared to the KORA-FF4 cohort. (c) The respective SVM coefficients found for KORA-FF4 and (d) L4L cohorts are similar for correlating medical conditions, such as diabetes and hypertension in the KORA-FF4 cohort and hypertension and heart disease in the L4L cohort (Figure 4.2c, d). Acronyms: SVM - support vector machine; NSP* - non-symptomatic individuals; AUC - area under the curve.

To achieve disease-specific classification outcomes it is, therefore, necessary to reduce the effect of correlating medical conditions and common parameters via matching cases and controls for them. In this study, cases and controls are matched for age, gender, BMI, smoking status and, if known, CRP concentrations as well as for comorbidities considering only the four medical conditions of interest in this section (see section 2.3.3 for more details about how matching is performed). The CRP values are mostly unknown for diabetes and asthma in the L4L cohort (Figure 3.13c), for which they are not matched. However, their characteristic signature is easy to address (section 3.2.2 and 3.2.3). Alcohol consumption is excluded from matching because partially unknown in the L4L study and because it might introduce unwanted biases being a self-reported qualitative factor [90, 91, 125, 126]. Tables 4.4 and 4.5 show the

distributions of common parameters for the cases of the four common conditions for KORA-FF4 and L4L cohorts after matching them to the respective controls. Similar values are found for the controls because of the matching. Since the matching algorithm tries to retain all cases, the distributions are similar to the ones reported in the previous chapter without matching (see Table 3.1 and 3.4). However, for heart disease and hypertension classes in the L4L cohort, the individuals with unknown CRP concentrations have been excluded from the analysis leading to slightly different distributions than in the classification with unmatched NSP* controls.

KORA-FF4 matched cohorts				
Cohort	n. cases	M/F	Age	BMI (kg/m^2)
Diabetes	288	1.5	69.4 ± 10.1	31.1 ± 5.4
Heart attack	69	3	71.1 ± 9.4	30.3 ± 5.5
Hypertension	1032	**1**	64.6 ± 11.3	29.3 ± 5.2
Asthma	181	**0.6**	**60.1 ± 12**	28.2 ± 5.6

Table 4.4: Distribution of common parameters among the cases of diabetes, heart attack, hypertension and asthma classes of KORA-FF4 cohort for the classification with matched controls. Acronyms: M/F - males-to-females ratio; BMI - body mass index.

L4L matched cohorts				
Cohort	n. cases	M/F	Age	BMI (kg/m^2)
Diabetes	30	1.7	67 ± 13.7	28.8 ± 5
Heart disease	27	2.9	72.3 ± 11.1	27.9 ± 6.1
Hypertension	48	**3**	69 ± 11.2	28.3 ± 5.8
Asthma	28	**1.1**	**53.6 ± 13.5**	27.5 ± 6.9

Table 4.5: Distribution of common parameters among the cases of diabetes, heart disease, hypertension and asthma classes of L4L cohort for the classification with matched controls. The values found for the common parameters are comparable with the ones observed for the KORA-FF4 cohort, with the exception of the M/F ratio for hypertension and asthma and of the average age for asthma. Acronyms: M/F - males-to-females ratio; BMI - body mass index.

Figure 4.4e and f show the SVM coefficients obtained for KORA-FF4 and L4L for the four medical conditions classified with unmatched NSP* controls (grey lines) and with matched controls (colored lines). The SVM coefficients obtained with the two analyses are very different, except for hypertension in the KORA-FF4 cohort for which the number of cases and controls are about 50% of the population and are almost the same after matching. The signatures of diabetes are also comparable in the two analysis despite the huge reduction in the number of controls in the classification with unmatched NSP* individuals to the one with matched controls (from 900 to about 300 individuals in the KORA-FF4 cohort and from 440 to about 25 in the L4L cohort.).

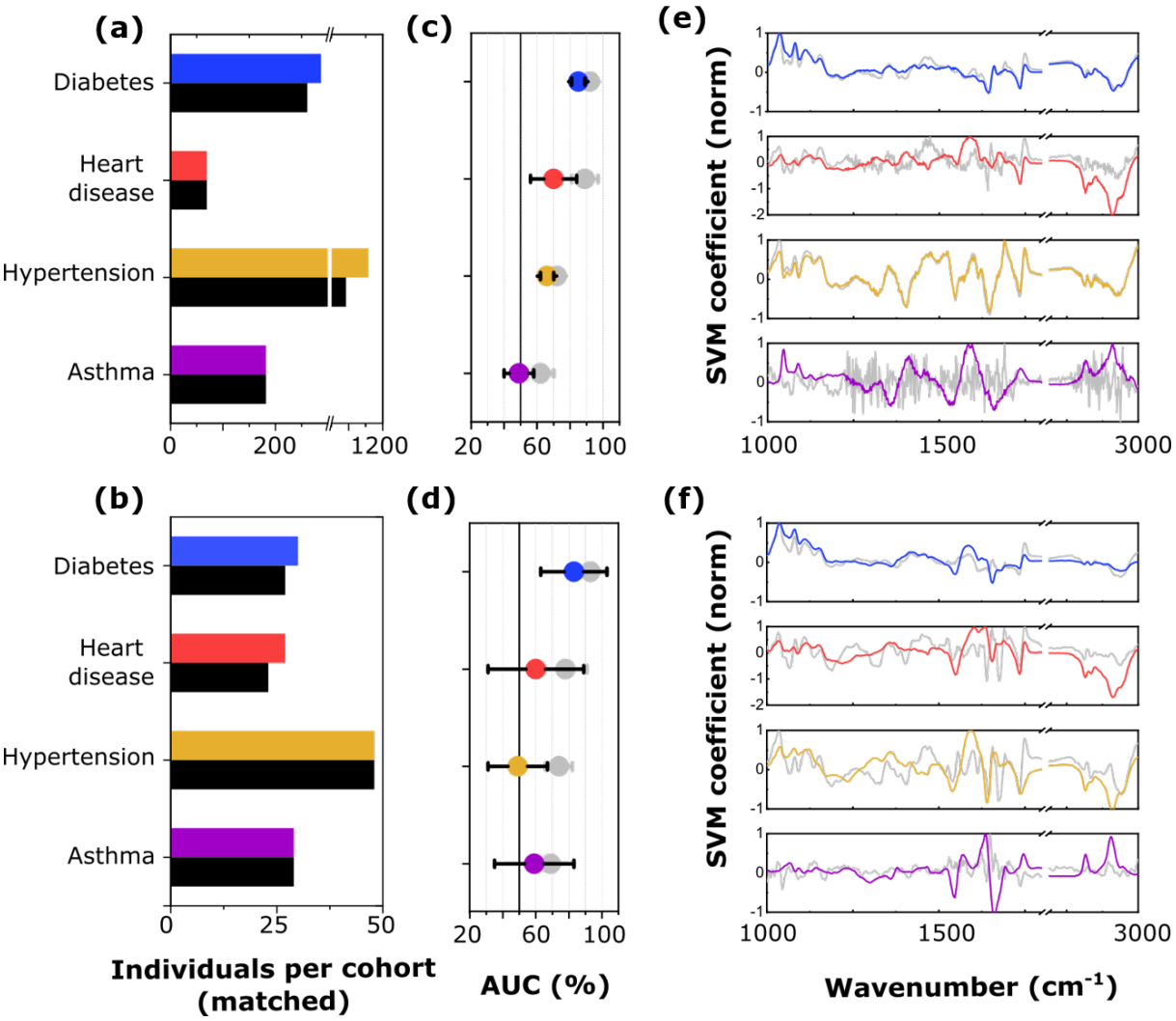

Figure 4.4: SVM binary classification of four common medical conditions with matched controls on the FT-IR spectra of KORA-FF4 and L4L cohorts. (a) The number of cases (colored) and controls (black) in KORA-FF4 and (b) L4L cohorts show that there are fewer cases for the same medical conditions in the L4L cohort. Because of the way matching is performed, the number of cases is the same as the number of controls. (c) The AUC scores for the four medical conditions are reported for KORA-FF4 and (d) L4L cohorts and have the same trend for both studies. The AUCs found from the classifications with matched controls (colored) are smaller and have higher standard deviations than the AUCs obtained in the classifications with NSP* unmatched individuals (grey) because matching cases and controls reduces the number of controls and, more importantly, reduces the contributions unspecific to the target medical condition due to parameters and comorbidities correlating with it. (e) The corresponding SVM coefficients normalized to their maximum are shown for KORA-FF4 and (f) L4L cohorts for the classifications with matched controls (colored lines) and with NSP* unmatched individuals (grey lines). The coefficients obtained for each medical condition with the two analysis are different and the contributions due to correlating comorbidities are reduced (Figure 4.5b) leading to disease-specific signatures, highly comparable for each medical condition between the two cohorts (Figure 4.5a). Acronyms: SVM - support vector machine; AUC - area under the curve; NSP* - non-symptomatic individuals.

In the classification with unmatched NSP* individuals, diabetes, hypertension and heart disease have a similar signature at low frequency (Figure 4.3c). However, the SVM coefficients obtained from the classifications with matched controls show that this signature is amplified for diabetes and reduced for heart disease and hypertension (Figure 4.4e), highlighting how it originated in the last two conditions from their correlation with diabetes (Figure 4.2c). This is shown in more detail for diabetes and hypertension in Figure 4.5b: these medical conditions influence each others´ SVM coefficient because of their correlation (p-value of $7.1 \cdot 10^{-25}$, Table 3.6). Similarly, the SVM coefficients found for diabetes, hypertension and heart disease have very comparable signatures if classified with unmatched NSP* controls because of their mutual correlation (Tables 3.6 and 4.3), while they return different coefficients if classified with matched controls. These observations prove that each parameter that affects the FT-IR fingerprints and correlates to any extend with the target phenotype affects its binary classification outcomes and it highlights that matching reduces the unwanted not-targeted contributions. Even more important is that the SVM coefficients found for the same medical condition are highly comparable between the KORA-FF4 and L4L independent cohorts (Figure 4.5a), thus unveiling that matching allows the identification of disease-specific signatures. The SVM coefficients found from the binary classification with matched controls for diabetes and asthma have a higher protein signature for the L4L cohort compared to the KORA-FF4 cohort which resembles the signature of inflammation (Figure 3.10b). The CRP concentration is, as explained above, not matched in the L4L cohort for these two medical conditions because it is unknown for most of the cases.

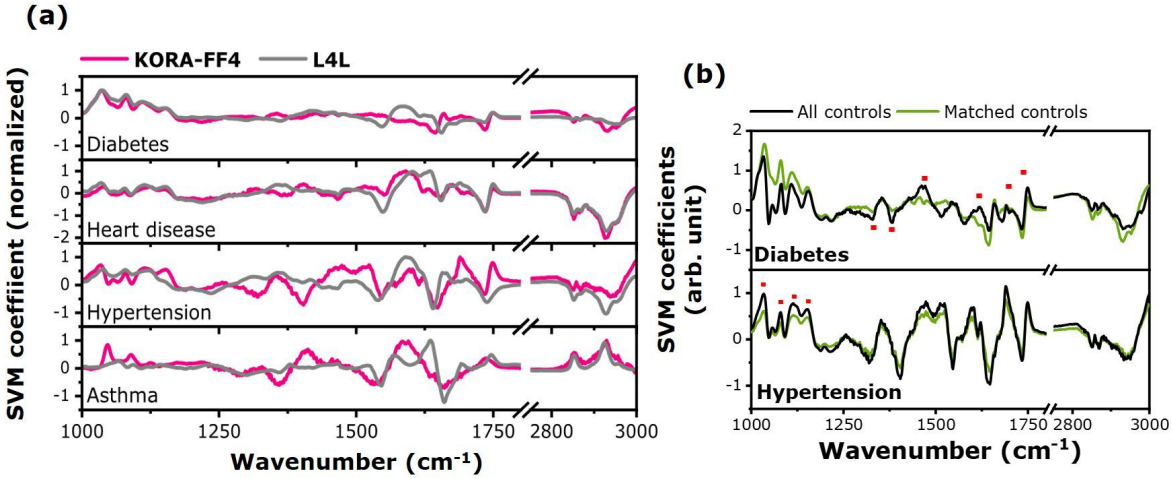

Figure 4.5: Disease-specific SVM coefficients of four medical conditions classified with matched controls for the FT-IR spectra of KORA-FF4 and L4L cohorts. (a) The SVM coefficients found from the classifications with matched controls return highly comparable signatures for the same medical conditions in the KORA-FF4 and L4L cohorts. Because of the very different nature of the two studies, these signatures are expected to be disease-specific. (b) The comparison between the SVM coefficients obtained from the classification of diabetes and hypertension in KORA-FF4 with NSP* individuals and with matched controls highlights that, because of their correlation (Table 4.2), these two medical conditions mutually affect their SVM coefficients unless cases and controls are matched for the not-targeted disease. The red dots highlight the features coming from their mutual interference, which are reduced after matching. Acronyms: SVM - support vector machine; NSP* - non-symptomatic individuals.

In conclusion, the comparison of two independent cohorts has allowed the identification of disease-specific IR signatures and classification efficiencies of four common medical conditions highlighting the importance of matching cases and controls to reduce the influence of correlating non-target factors. In particular, IR spectroscopy achieves the highest efficiency for the detection of diabetes, followed by heart disease and hypertension. However, according to this analysis, FT-IR spectroscopy does not appear suited for the detection of asthma for which the AUC is 50%. The higher AUC found in the classification of asthma with unmatched controls might be due to other factors correlating with asthma. Matching is therefore essential to address if the target condition leaves a trace on the IR fingerprint of human blood biofluids.

Several published studies aim at identifying the best biofluid between human blood plasma, used in the KORA-FF4 cohort, and serum, used in the L4L cohort, for disease detection via FT-IR spectroscopy [114, 127]. This study does not find any evident advantages of one biofluid over the other for the classification of the four conditions considered. This is due to the highly comparable chemical composition of the two blood biofluids [128]. Tables 4.4 and 4.5 show that the distribution of common parameters is comparable for each condition in KORA-FF4 and L4L, except for the males-to-females ratio for hypertension and asthma classes and the average age for asthma. While this is not relevant for the analysis of asthma which does not have a detectable IR signature in any case, the different gender distribution among the individuals affected by hypertension of the KORA-FF4 and L4L cohorts might compromise the veracity of the comparison between the two cohorts. This study is a first step needed to stress the importance of matching cases and controls for the identification of disease-specific signatures via the analysis of very different independent cohorts. However, more extensive studies on larges cohorts and more medical conditions are necessary to establish that these conclusions are generally valid and would permit the definition of standard procedures for how to match cases and controls for each medical condition.

The identified disease-specific spectral signatures highlight that the features associated with carbohydrates and proteins' glycosylations (1000 – 1250 cm^{-1}, Table 2.1) are responsible for the detection of diabetes, in agreement with the well-known role of blood glucose in diabetic patients, as well as with the effect of high concentration of blood glucose on the protein glycosylation, such as for hemoglobin and albumin, which are increasingly used to monitor and detect diabetes [129, 130] because of the high accuracy of the glycemic control over a long period that these biomolecules offer. The classification efficiency found for diabetes in this study is in agreement with other studies reported in the literature based on different infrared spectroscopic techniques [131] or biofluids [25]. In particular, [131] reports the ATR-FTIR analysis of human whole blood combined with XGBoost algorithms and reports a sensitivity of about 95% on a very small number of cases (50 individuals), while the sensitivity found in this study is about 90% for the classification of about 300 type II diabetes cases. Lipids (1750 – 3000 cm^{-1}) return the most relevant spectral signatures for the detection of heart disease, in agreement with the acknowledged connection between an increased blood serum concentration of low-density lipoprotein (LDL) with the higher risk of experiencing an episode of heart attack [132, 133]. The comparison between the FT-IR spectra of HDL and LDL molecules with the spectrum of human blood serum reported in the literature highlights that the signature of heart disease found in this study is indeed related to these biomolecules [134]. The smaller signature at lower frequency due to carbohydrates can be explained by the observation reported in the literature according to which individuals who experienced a heart attack tend to have

a lower fasting blood glucose concentration than the respective controls [135]. The spectral features underneath the classification of hypertension are spread in the whole spectrum and the interpretation is more difficult. It has been reported in the literature that the concentration of urea and creatinine in human blood serum and saliva are potential biomarkers for hypertension [136], both with spectral features between 1000 and 1800 cm^{-1}. However, further analyses are necessary to establish the nature of the fingerprint of hypertension observed in this study. To the best f our knowledge, this is the first study addressing the FT-IR signature of hypertension in human blood serum and plasma. Diabetes, hypertension and cardiovascular diseases, in general, are connected with lifestyle and diet. A diet rich in cholesterol and saturated fatty acids, indeed, has been reported to reduce the LDL receptors increasing the LDL blood concentration [137]. Moreover, lifestyle factors such as diet and little physical activity promote an increase in BMI which is connected with conditions like insulin resistance and diabetes [138], while the high consumption of sodium or alcohol tend to increase the blood pressure and can induce hypertension [139]. The fast and early diagnosis of these conditions can help many individuals modifying their lifestyle timely thus preventing them from contracting these medical conditions. In this context, FT-IR can be a powerful method for the fast, early and simultaneous diagnosis and monitoring of these diseases.

The drawback of matching is the reduction in the number of controls which, together with the effect of reducing the non-target contributions, returns lower AUCs with higher standard deviations (Figure 4.4c, d). However, the AUC obtained matching cases and controls are more disease-specific. Moreover, equalizing the number of cases and controls increases the features in the SVM coefficient. This is the case of heart disease and asthma which count a low number of individuals: in the classifications, with all 900 NSP* individuals for the KORA-FF4 study, these cohorts return almost featureless SVM coefficients, while if classified with a comparable number of controls the SVM coefficients have strong signatures (Figure 4.4e). The number of cases impacts the classification outcomes. Before addressing this is in section 4.2.3, the FT-IR fingerprints and classification efficiencies of other common medical conditions known for KORA-FF4 are addressed.

4.2.2 FT-IR fingerprinting of common medical conditions in KORA-FF4

The previous section has focused on the analysis of four medical conditions for the comparison of the KORA-FF4 and the L4L cohorts. However, more medical conditions are known for the KORA-FF4 cohort (Table 3.1). This section reports the SVM binary classifications of the FT-IR spectra of all endpoint medical conditions known for the KORA-FF4 study, in particular for individuals affected by type II diabetes (referred to as "diabetes"), hypertension, asthma, high blood lipids, chronic obstructive pulmonary disease (COPD), as well as for individuals who had cancer (former or ex-cancer cohort) or experienced episodes of stroke or heart attack (previously referred to as "heart disease").

Figure 4.6a shows the number of individuals for each medical condition. Being a prospective cross-sectional population-based cohort, some medical conditions are underrepresented since only a few individuals have developed them during the study. The individuals symptomatic for hypertension and high blood lipids constitute the classes with more cases (about 50% of the whole cohort), followed by diabetes, former cancer patients and asthma, while all the

other medical conditions count less than 150 cases each. The classification outcomes are affected by the number of cases, as addressed in the next section. The AUCs highlight the higher potential of FT-IR spectroscopy on human blood plasma for the detection of diabetes, hypertension and high blood lipids (Figure 4.6b), which, incidentally, have the largest number of cases. The classifications with all non-symptomatic individuals (NSP/NGT, analyzed in section 3.2.2) are compared with the classifications with matched controls. As previously shown for diabetes and hypertension (Figure 4.5b), matching cases and controls reduces the impact of not-targeted factors and, therefore, returns lower AUCs compared to the classification with unmatched controls (Figure 4.6b). In particular, cases and controls have been matched for the common parameters age, gender and CRP being these the most important factors affecting the between-person spectral variability in the whole spectral range (section 3.2.2), as well as for all comorbidities with high AUCs in the classification with all NSP/NGT individuals, namely diabetes, hypertension and high blood lipids, and the ones that correlate with the target condition (p-value > 0.01), as it is the case for asthma and COPD.

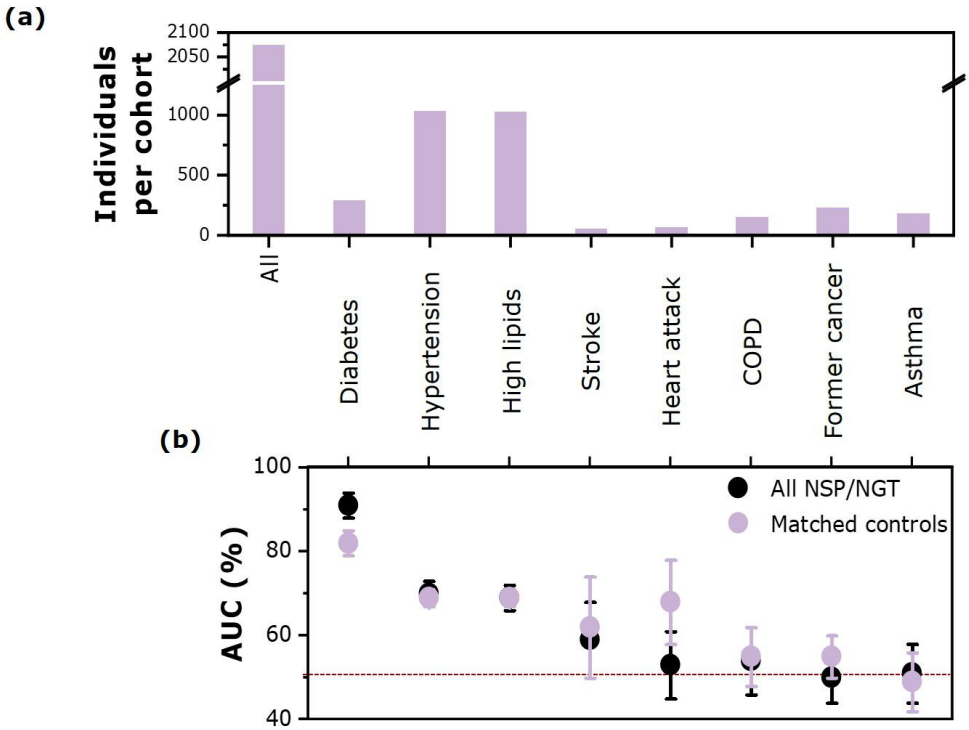

Figure 4.6: SVM binary classification of all known endpoint medical conditions with all NSP/NGT unmatched individuals and with matched controls on the FT-IR spectra of KORA-FF4 cohort. (a) The number of individuals for each medical condition is reported for the KORA-FF4 population (Table 3.1) and reflects the distribution of each disease in the represented general population. (b) The comparison between the AUCs derived by the classifications of each disease with all NSP/NGT unmatched individuals (black) and with matched controls (light purple) shows that matching cases and controls reduces the number of controls and the non-target features, thus isolating the disease-specific fingerprint and, therefore, resulting in lower AUCs with larger standard deviations. Acronyms: SVM - support vector machine; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals; AUC - area under the curve; COPD - chronic obstructive pulmonary disease.

Section 3.2.2 addresses the main sources of the between-person spectral variability among healthy, mainly lipids and proteins influenced by age and inflammation. Because of the large between-person variability of proteins and lipids and of their spectral signatures (1750 - 3000 cm^{-1} and 1250 - 1750 cm^{-1} respectively, Table 2.1), the differential fingerprints of the different medical condition are similar to each other (Figure 4.7) and have similar standard deviations for cases (colored shaded areas) and controls (black shaded areas), especially between 1250 and 3000 cm^{-1}. Moreover, the strong protein features from the Amide bands, which are in common with all differential fingerprints, are also affected by the largest technical noise among all spectral features (Figure 3.8b, c).

Despite the strong between-person variability, the SVM algorithm identifies features for each condition. In particular, the SVM coefficients shown in Figure 4.7 highlight the features responsible for the binary classification with the NSP/NGT unmatched individuals (black line) and with matched controls (white lines). Both hypertension and high blood lipids return similar AUCs and SVM coefficients for both classifications because the matching algorithm retains about all cases and controls. A deeper inspection of the SVM coefficient of hypertension unveils the reduced intensity of the features at lower frequencies after matching the distribution of diabetes between cases and controls, as already shown in Figure 4.5c. Similarly, the SVM coefficient found for diabetes has higher signatures at low frequency in the classification with matched controls compared to the unmatched ones. The SVM coefficients found for high blood lipids in the two classifications are the same because this condition does not strongly correlate with any other known medical condition or common parameter (Figure 4.2c). The comparison of the SVM coefficients of the other medical conditions is more difficult because the high discrepancy in the number of cases and controls for the classification with all NSP/NGT individuals return featureless SVM coefficients (section 4.2.3). After matching, the SVM coefficients show similar features as the differential fingerprints, with stronger differences in the Amide I and II bands, probably because of the larger technical noise recorded for this spectral region (Figure 3.8b, c).

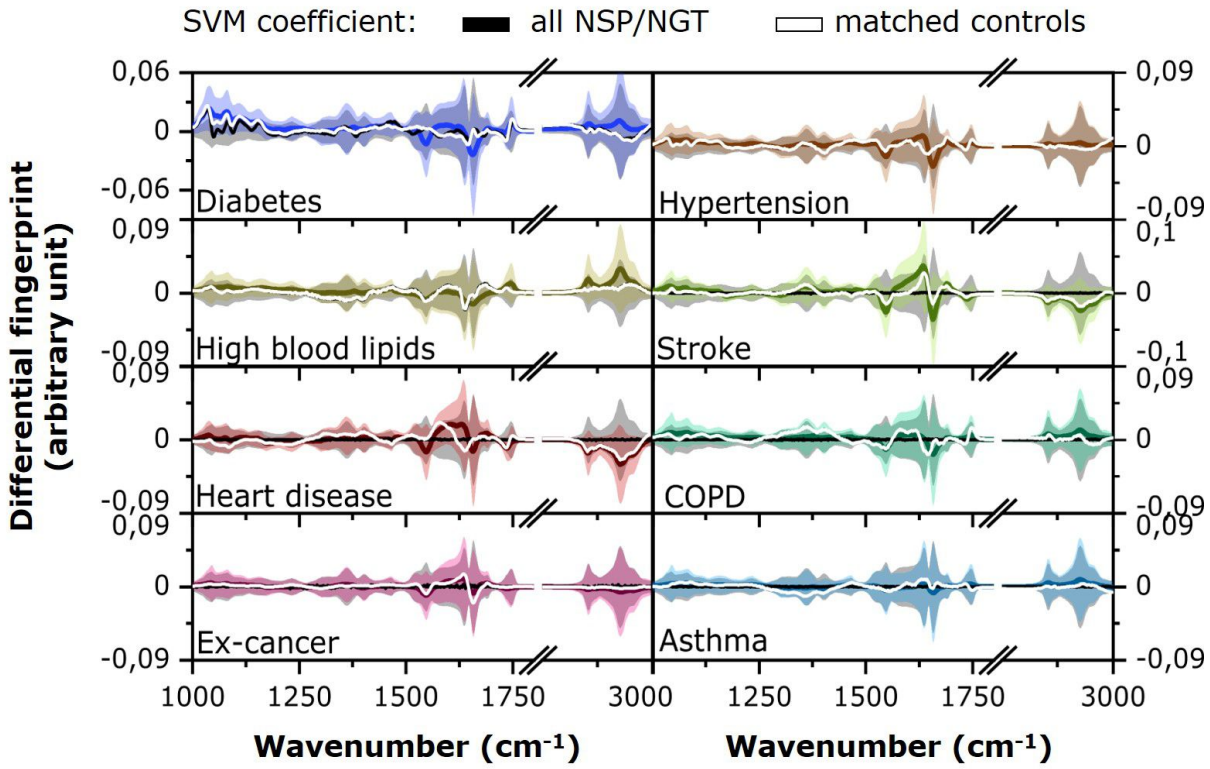

Figure 4.7: Differential fingerprint and SVM coefficients of all known endpoint medical conditions classified with all NSP/NGT unmatched individuals and with matched controls on the FT-IR spectra of KORA-FF4 cohort. The colored lines are the differential fingerprint of the respective medical condition, calculated as the difference between the average spectrum of cases and the average spectrum of all individuals non-symptomatic for that condition. The colored shaded area is the standard deviation of the cases and the black shaded area is the standard deviation of the controls. The standard deviations of cases and controls are comparable above 1250 cm^{-1} because of the large between-person spectral variability of proteins and lipids (section 3.2.2). The corresponding SVM coefficients (scaled) are reported for comparison both for the classification with all NSP/NGT individuals (black line) and with matched controls (white line). The SVM coefficients show that, despite the large between-person spectral variability, the SVM algorithm identifies disease-specific signatures. The comparison between the classification with unmatched and with matched controls highlights again that matching reduced the non-target features due to factors correlating with the target-disease, as seen in the previous section. Acronyms: SVM - support vector machine; NSP/NGT - cohort of non-symptomatic normal glucose tolerance individuals.

In conclusion, the large spectral variability analyzed in section 3.2.2 affects the variability recorded in the differential fingerprints of all medical conditions, but it does not prevent SVM from identifying disease-specific signatures for all medical conditions. The signature of diabetes, hypertension and heart attack classes have been discussed in the previous section. The other common parameters leading to high classification efficiencies are high blood lipids and stroke. The features underneath the classification of high blood lipids are due to protein and lipids, probably due to the vibrations of lipoproteins [134], while the smaller contributions at lower frequencies could be due to triglycerides and glycerol [140]. The spectral signatures responsible for the classification of stroke are also spread on the whole spectral range, with the

main contributions from proteins and lipids. The risk factors identified for stroke are multiple, among which high plasma levels of lipoproteins [141] which could explain the observed spectral signature. The plasma concentration of several amino acids [142] and of organic and inorganic metal complexes [143] have also been proposed as biomarkers for stroke and could explain the signatures at lower frequencies.

The highest classification efficiencies are found for the most numerous classes, namely diabetes, hypertension and high blood lipids. Despite the small number of cases, the AUCs are promising also for stroke and heart attack, even if the standard deviations are large. The analysis of larger cohorts of individuals who experienced episodes of stroke or heart attack could help to define the signatures specific to these conditions to identify people at high risk. COPD, former cancer and asthma classes return AUCs close to 50%. However, a negative outcome can be considered true only for cohorts large enough to guarantee a good statistical power of the study [144], as explored in the next section.

4.2.3 Influence of the cohort size on binary classifications

The medical conditions with the highest AUCs are found to be also the most numerous classes (Figure 4.6b). The analyses of the classes with a few cases have a low statistical power, which is the probability of detecting a specified clinical significance (e.g. the probability to obtain AUC > 50%) [144–146]. The minimum number of cases necessary has been frequently discussed in pattern recognition [147–149], but defining the smallest cohort to achieve a reliable classification efficiency in studies similar to the one presented here is not straightforward [150]. The FT-IR spectra recorded for the large population-based cohort of KORA-FF4 are useful for this purpose. Achieving robust classifications is essential for an objective comparison of the FT-IR performances for the diagnosis of different medical conditions. Generally, higher AUCs correspond to stronger IR signatures, which are therefore easier to identify and have a higher significance. Since the AUC depends only on the strength of the signal and not on the features pattern, the analysis of medical conditions with the same AUCs are expected to have the same statistical power and to require the same minimum number of cases to achieve robust classifications. Diabetes and hypertension are selected for this analysis because these are two of the largest classes in the KORA-FF4 cohort and because they have very different AUCs. Diabetes is expected to be representative of all conditions with AUC close to 85% in the classification with matched controls while hypertension is expected to represent all conditions with AUC close to 66%. All medical conditions with intermediate AUCs are expected to require an intermediate minimum number of cases for roust classifications.

One way generally used to estimate the required sample sizes for medical classifications is using learning curves [151, 152], graphical representations of the classification efficiency for an increasing number of cases. The learning curves for the AUCs obtained via SVM classifications show that the classification efficiencies increase for the increasing number of cases (Figure 4.8a, top and medium panels). The learning curves are reported for the classifications of the two medical conditions with all the unmatched NSP* controls (defined in section 3.2.3; black lines) as well as with matched controls (colored lines). All the classifications with the NSP* individuals are performed keeping the same 900 individuals as controls, while in the classifications with matched controls the number of controls is always the same as the cases.

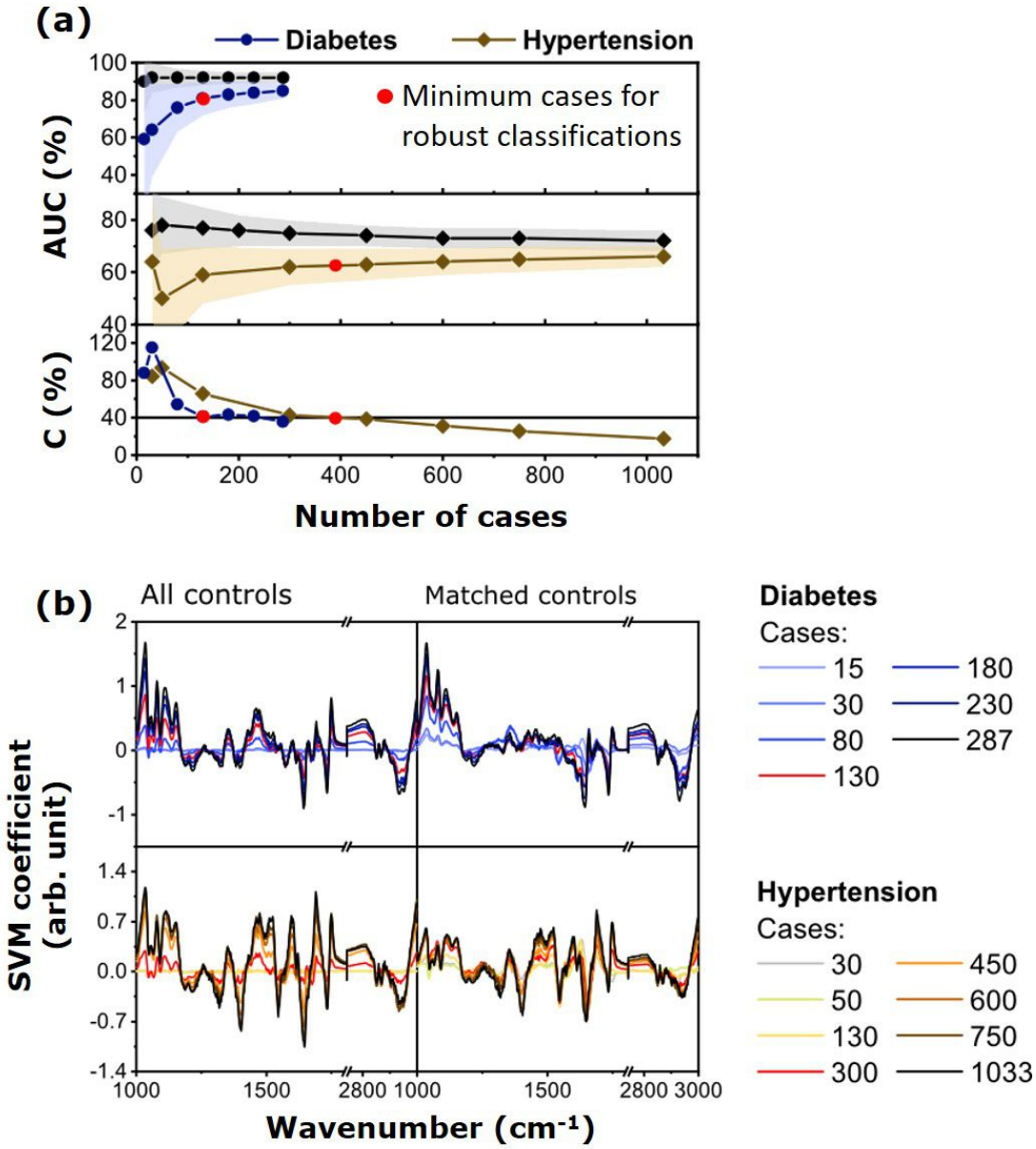

Figure 4.8: AUC and SVM coefficients for an increasing number of cases from the classification of diabetes and hypertension with all NSP* unmatched individuals and with matched controls on the FT-IR spectra of KORA-FF4 cohort. (a) The learning curves show the dependence of the AUC to the number of cases for the classifications with all NSP* individuals (black lines) and with matched controls (colored lines) for diabetes (top panel) and hypertension (middle panel). The AUCs found with matched cases and controls are always significantly lower than with unmatched controls. The bottom panel shows the relative difference of the respective SVM coefficients C defined in the text. This panel shows that the coefficients found in the classifications with matched and with unmatched controls are significantly different and never zero. In general panel (a) highlights that matching reduces the non-target features of factors correlating with the target condition leading to lower AUCs and giving disease-specific signatures. The red dots identify the smallest number of cases for robust classification, namely corresponding with an AUC 5% lower than the AUC at saturation. (b) The SVM coefficients corresponding to the top and middle panels of (a) show that the smallest number of cases for robust classifications are also the smallest cohort to return the same SVM coefficients (red lines) as for larger cohorts. Acronyms: AUC - area under the curve; SVM - support vector machine; NSP* - non-symptomatic individuals.

The SVM coefficients found for each medical condition are similar in both analyses for large cohorts (Figure 4.8b). However, for small cohorts, the SVM coefficients found with matched controls have slightly different features than for a larger number of cases. If the number of cases is too small compared to the number of controls, the SVM coefficients do not have evident features despite the high classification efficiencies. This is the case for the classifications of less than 100 cases against the 900 NSP* individuals for which the AUCs reach high values. The AUCs from the classification of more cases with the NSP* individuals reach a plateau for diabetes, while they slowly drop for hypertension. On the contrary, the AUCs found for the classifications with matched controls rise for an increasing number of cases until saturation. Fitting the learning curves with logistic functions returns a maximum AUC of 85,0 ± 0,2% for diabetes and 65,9 ± 0,8% for hypertension. For higher AUCs, the disease detection is easier and the significance of the feature is higher, which implies that a given statistical power can be reached with smaller cohorts [145]. This explains the faster exponential growth found for diabetes compared to hypertension (Figure 4.8a).

The smallest number of cases to achieve robust classifications is the one giving an AUC as close as possible to saturation as well as stable SVM coefficients. In this study, the minimum number of cases required for robust classification is defined as the one for which the AUC is 5% lower than the AUC at saturation. The minimum number of cases required is 130 cases for diabetes, giving an AUC of 81%, and 380 cases for hypertension, leading to an AUC of 63% (Figure 4.8a, red dots). Incidentally, the identified number of cases are also the smallest classes for which the SVM coefficients have the same features as for larger cohorts for both diabetes and hypertension (Figure 4.8b). The SVM coefficients in the classifications with NSP* are different from the SVM coefficients found for the classifications with matched controls. Their relative difference can be easily quantified in a coefficient C for any given number of cases n. Since the SVM coefficient obtained for the classifications with all NSP* individuals do not have evident features for a too small number of cases, the coefficient obtained for the maximum number of cases available (N) is taken as reference for each medical condition. The relative difference C is therefore here defined as:

$$C(n) = 100 * \frac{\sum_{\omega} |(S_m(n,\omega) - S(N,\omega)|}{\sum_{\omega} |(S(N,\omega))|} \tag{4.1}$$

where N is 288 for diabetes and 1035 for hypertension (Table 3.1), S_m are the SVM coefficients found for the classifications with matched controls normalized to their maximum, S are the SVM coefficients found for the classifications with NSP* individuals and ω is the spectral coverage of FT-IR excluding the silent region between 1800 and 2750 cm^{-1}. For both conditions, the difference C found for the minimum number of cases for robust classifications is about 40% (Figure 4.8a, lower panel - red dots). In other words, the SVM coefficients found for the classification of the smallest number of cases necessary with matched controls are significantly different from the ones obtained for the largest cohorts in the classifications with all NSP* individuals. This proves that, as concluded in the previous sections, matching significantly reduces the non-specific features present in the SVM coefficients found for the classifications with unmatched control (Figure 4.5). This is again evident from the lower AUC found for the classifications with matched controls. Even though the AUCs obtained for hypertension in the classifications with unmatched and with matched controls seem to converge (Figure 4.8, middle panel), fitting the two trends with logistic curves reveal that the AUCs at the respective

plateaus are significantly different.

The results found for diabetes and hypertension can be generalized under the realistic assumption that only the magnitude of the signature affects the statistical power of the study independently on the specific signatures. Therefore, it can be concluded that the minimum number of cases required for robust classifications is expected to be below 130 cases for AUCs higher than 85%, as for diabetes, and more than 380 cases for AUCs lower than 66%, as for hypertension. In light of these observations, a proper comparison of the classification efficiencies obtained for different medical conditions should be done for the number of cases equal to or larger than the corresponding minimum required for robust classifications with matched controls. However, except for the cohort of individuals with high blood lipids, the other conditions count too few cases. Therefore, the best comparison possible is for the same number of cases. The AUC corresponding to 70 cases is about 74% for diabetes, 70% for heart attack and 55% for hypertension. According to these values, the maximum AUC for heart attack is expected to be close to the one found for diabetes. Similarly, the AUCs found for the classification of 182 asthma and hypertension cases are respectively 50% and 60%. Therefore, for a robust classification, more than 380 asthma cases are expected to be necessary to address whether the FT-IR spectra of human blood plasma are suited for the detection of this condition. However, whether a poor outcome is worth spending voluminous resources is not trivial [145]. Similar conclusions can be drawn for the 230 former cancer patients and the 150 COPD cases for which the AUCs are smaller than for the same number of cases with hypertension. On contrary, for the stroke class which counts only 54 cases, the AUC found is about 60% and it would be worth addressing the best classification efficiency in larger cohorts to better address the potential of FT-IR for the screening individuals at high risk.

In summary, the learning curves show significantly different trends for the classifications with all unmatched NSP* individuals compared to the matched controls because matching cases and controls removes non-target contributions returning disease-specific SVM coefficients and systematically lower classification efficiencies. The classifications with matched controls are therefore used for the comparison of the classification efficiencies of each medical conditions analyzed for the same number of cases, from which it can be concluded that a very large number of cases is required for the robust classification of COPD, asthma and former cancer patients for which the classification efficiencies are expected to be lower than the 65% found for hypertension. This suggests that the SVM analysis of the FT-IR spectra of human blood plasma might not be suited for the classification of these medical conditions. On the other hand, the classification of individuals who had episodes of heart attack or stroke return promising efficiencies for a small number of cases, highlighting how the analysis of larger cohorts might lead to the possibility of timely identifying features characteristic of individuals at high risk. Diabetes, hypertension and high blood lipids classes count enough cases for robust SVM classifications. In conclusion, the study reported here shows that the minimum number of cases required to address the efficiency of SVM binary classifications of FT-IR human blood plasma fingerprints is 130 as observed for diabetes, which is comparable to the conclusions reported in the literature for another classification algorithm for the Raman spectra of single cells [150] which identifies 75-100 cases as the minimum number to achieve good but not perfect classifications. These conclusions might be applied to similar studies to address the reliability of AUCs of 50%.

4.2.4 Comparison of FRS and FT-IR fingerprints of common conditions in KORA-FF4

The potential of the newly developed FRS spectroscopy for the diagnosis of diseases on human blood plasma is investigated for all known medical conditions of the KORA-FF4 cohort (Table 3.1) and compared with the FT-IR spectra in the same spectral range (1000 and 1500 cm^{-1}).

The SVM classifications of the FT-IR spectra in the reduced an in the full-spectral coverage (section 4.2.2) are comparable (Figure 4.9a). For both FRS and FT-IR data, the classifications of each medical condition with all individuals non-symptomatic for the target disease ("unmatched controls") return higher or comparable AUCs than for the classifications with matched controls (Figure 4.9), in agreement with what concluded in section 4.2.3.

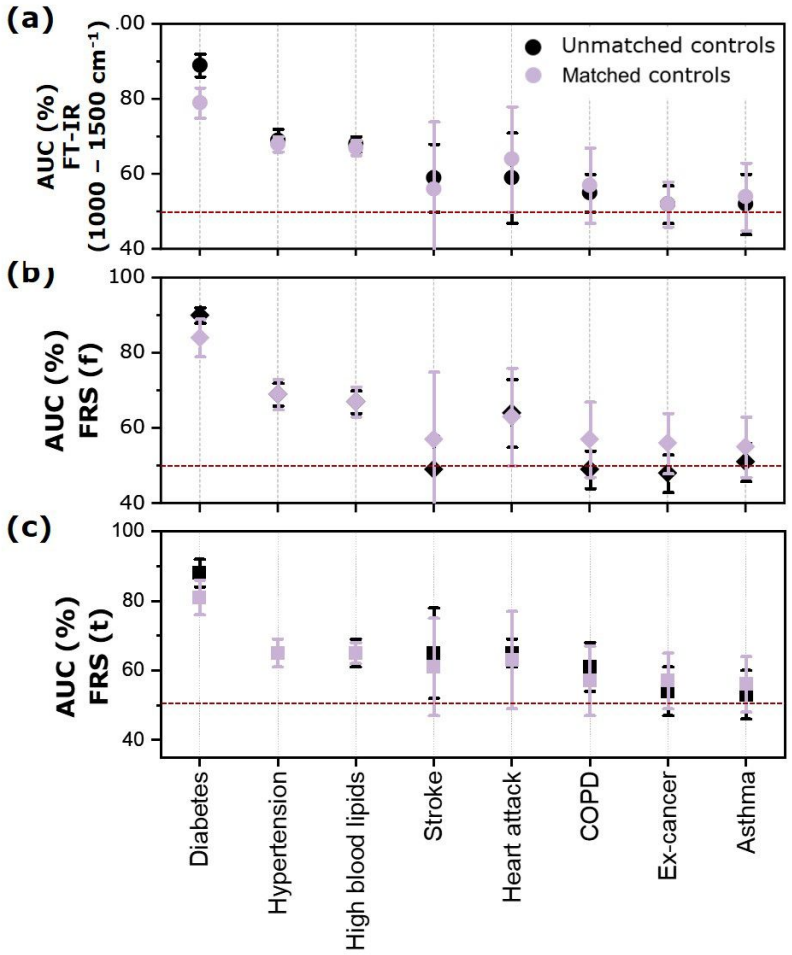

Figure 4.9: SVM classification efficiency of all endpoint medical conditions with matched and unmatched controls on the FT-IR ad FRS data of KORA-FF4 cohort. (a) The AUCs are reported for FT-IR reduced spectra (1000 - 1500 cm^{-1}) and for the corresponding (b) FRS data in the frequency domain (*FRS(f)*) and (c) in the time domain (*FRS(t)*). The classifications with all individuals non-symptomatic to the target medical condition ("unmatched controls") return higher or comparable AUCs that with matched controls. Acronyms: SVM - support vector machine; FRS - field-resolved spectroscopy; AUC - area under the curve; COPD - chronic obstructive pulmonary disease.

The FRS data in the frequency domain are calculated as "absorbance" after applying the optimal HTPF from 0.14 to 7 *ps* (section 3.1.3). As explained in section 2.3.2, the SVM classifications of the EMFs are performed by scanning time windows with different widths to identify the *optimal time window*, namely the temporal range giving the highest AUC. The classification efficiencies found for FRS are comparable with the ones of FT-IR in the same spectral range (Figure 4.10a). Given the large day-to-day dependence affecting the first PCs of FSR data both in the time and in the frequency domain (section 3.2.4), this is a promising first result. The SVM coefficients found for the FRS data in the frequency domain are noisy, but their absolute values have features in common with the SVM coefficients found for the FT-IR spectra (Figure 4.10b).

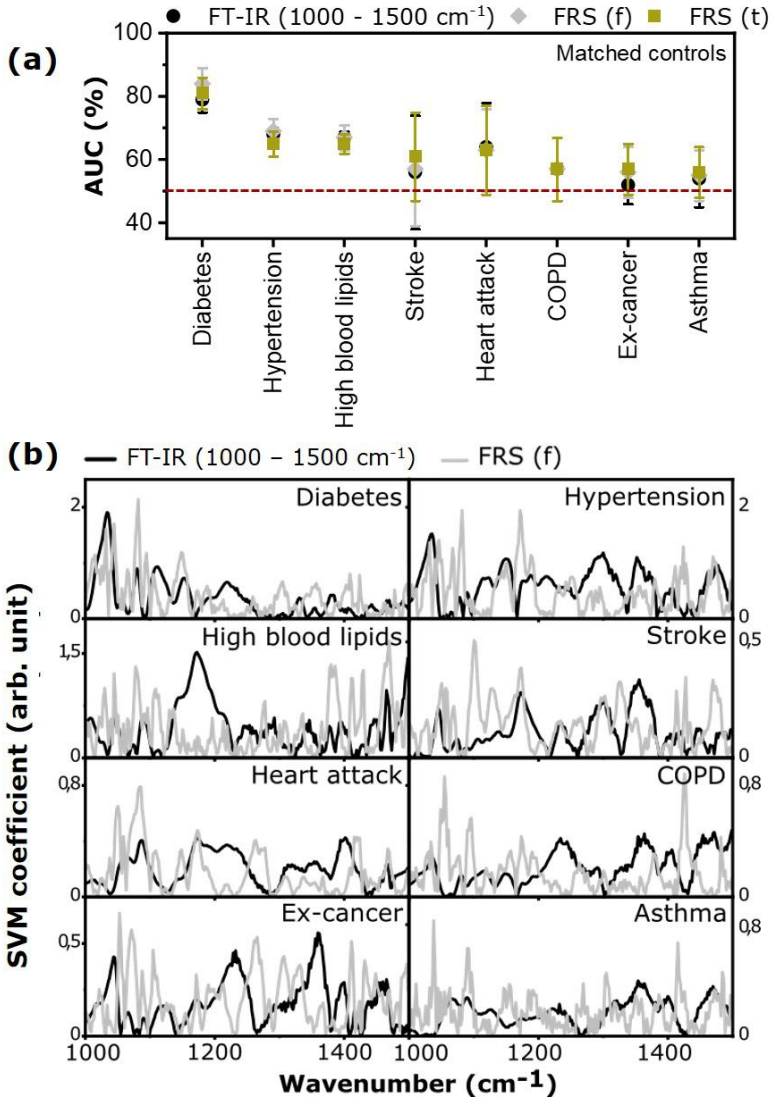

Figure 4.10: SVM classifications of all endpoint medical conditions with matched controls on the FT-IR ad FRS data of KORA-FF4 cohort. (a) The AUCs obtained via FRS and FT-IR between 1000 and 1500 cm^{-1} are comparable. (b) The absolute values of the corresponding SVM coefficients found for FRS in the frequency domain have common features with the coefficients found for FT-IR. Acronyms: SVM - support vector machine; FRS - field-resolved spectroscopy; AUC - area under the curve; COPD - chronic obstructive pulmonary disease.

The SVM coefficients obtained from the EMFs in the time domain are reported for the identified optimal time window (Figure 4.11). As observed for FT-IR in full-spectra, hypertension and high blood lipids return the same SVM coefficients for the classifications with matched and unmatched controls. This happens because these conditions affect about 50% of the population and the matching algorithm, which tries to retain as many cases as possible, leads to similar case/controls cohorts as without matching. However, as previously shown for FT-IR (Figure 4.5b), the SVM coefficients are similar but not identical because matching reduces the spurious contributions.

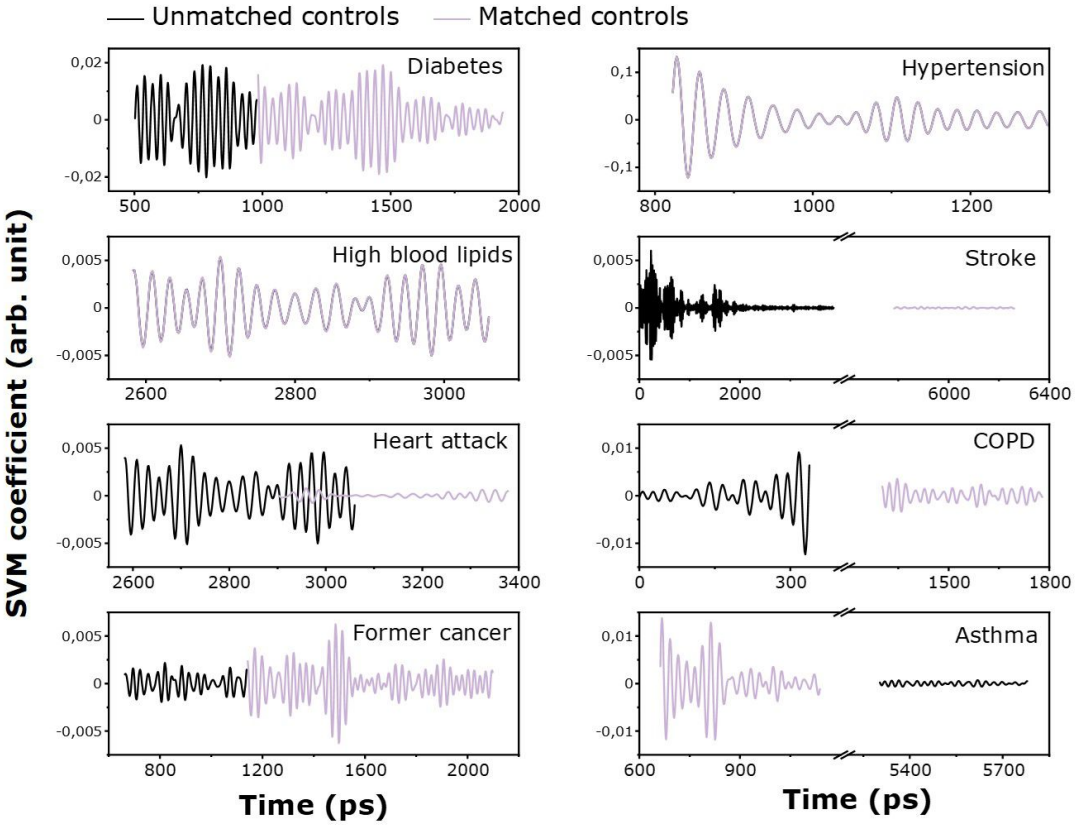

Figure 4.11: SVM coefficients of all endpoint medical conditions with matched and unmatched controls on the EMFs of KORA-FF4 cohort. The SVM coefficients are reported in the respective identified optimal time windows, which are different for the classifications with matched and unmatched controls for all medical conditions. The only exception are hypertension and high blood lipids for which, since they affect about 50% of the population, almost the same cases and controls are considered in the classifications with and without matching. Acronyms: SVM - support vector machine; EMF - electric-field-resolved molecular fingerprint; COPD - chronic obstructive pulmonary disease.

Except for hypertension and high blood lipids, the optimal time windows identified for the other medical conditions are different for the classifications with unmatched controls compared to the classification with matched controls (Figure 4.11). The time resolution can be advantageous to disentangle the contributions not specific to the target disease present in the classifications with unmatched controls and reduced by matching cases and controls. Four

showcases are considered (Figure 4.12): diabetes and high blood lipids, which count more cases than the minimum required for robust classifications (section 4.2.3) and result in very different AUCs, as well as heart attack and former cancer classes, which count fewer cases than the estimated minimum for robust classifications and return different AUCs.

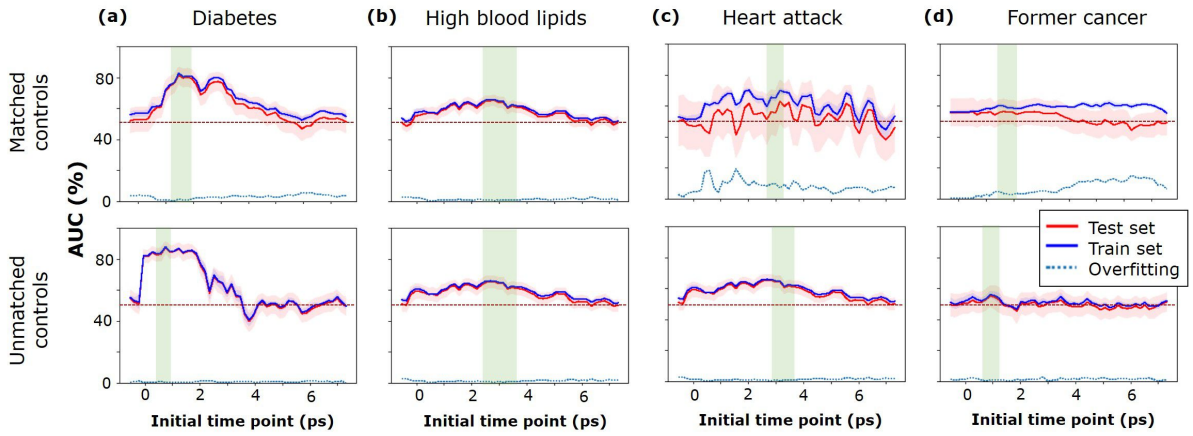

Figure 4.12: AUC trends along the time axis of four medical conditions with matched and unmatched controls on the EMFs of KORA-FF4 cohort. (a) the AUCs trends along the time axis are reported for diabetes, (b) high blood lipids, (c) heart attack and (d) former cancer patients for the classifications with matched (upper panels) and unmatched (bottom panels) controls. The red dashed lines show the AUC = 50%. The AUC can have values close to the maximum reached for that classification also outside the optimal time window (green shaded areas). In the classifications with unmatched controls, the AUCs have comparable values in the same optimal temporal windows found for the classifications with matched controls. However, the AUCs are higher for unmatched controls around 0 *ps* due to the non-specific contributions of factors correlating with the target medical condition. Acronyms: AUC - area under the curve; EMF - electric-field resolved molecular fingerprint.

The AUC trends along the time axis for these four medical conditions unveils that the AUCs are close to the maximum value reached for the corresponding classification for longer times than the identified optimal temporal windows (Figure 4.12, green shaded areas), which are strictly around the highest AUCs. After matching, the number of controls is reduced introducing larger fluctuations and overfitting rates, with a more important effect for the smallest cohorts such as heart attack and former cancer patients. The overfitting rates are calculated as the distance between the average AUC of the training set and the standard deviation of the test set (section 2.3.2). The class of individuals with high blood lipids offers the advantage to compare the classifications with matched and unmatched controls without the side effect of the extensive reduction in the number of controls due to matching. The AUCs trends found for high blood lipids are identical for the classifications with matched and unmatched controls (Figure 4.12b), with the only exception at shorter times (0 - 1 *ps*) for which the AUCs are higher in the classifications with unmatched controls. Similar conclusions can be obtained for the other conditions: in the classification with unmatched controls, the AUCs are close to the maximum between 0 and 2 *ps* for diabetes, which therefore covers also the optimal time window found for the classification with matched controls, between 0 and 3.5 *ps* for heart attack and between 0.5 and 1 *ps* for former cancer patients. However, the AUCs trends have

big fluctuations for the classification of small cohorts, such as heart attack and former cancer patients, with the matched controls, thus making the comparison with the classifications with unmatched controls difficult.

In summary, for all medical conditions, the AUC found from the classifications with all individuals non-symptomatic to the target condition (the "unmatched controls") reach the highest value around 0 *ps* and remains high for a period that covers the optimal time window found for the classification with the matched controls. In other words, thanks to the resolution in time it is possible to address that the classification with unmatched controls identify the disease-specific temporal signatures, isolated in the classifications with matched controls, plus the additional contributions of non-specific signatures due to factors correlating to the target disease which affect the initial times, where the signal is more intense. The time resolution offers an important advantage on frequency-resolved techniques because it separates disease-specific and unspecific signatures in different time windows. The classification efficiencies of all medical conditions investigated are comparable for FRS in both the time and frequency domain and FT-IR, which is a promising achievement considering the day-to-day dependence of the newly developed technique.

Despite matching isolates the disease-specific signatures, it reduces the number of controls leading to higher overfitting rates, detrimental especially for the medical conditions affecting a few cases. Therefore, larger studies are necessary to achieve a deeper understanding of the medical conditions underrepresented in the KORA-FF4 cohort for which FT-IR and FRS fingerprints are expected to be good predictors, such as stroke and heart attack (see section 4.2.3).

4.3 IR spectral liquid biopsy of an intermediate condition: prediabetes

The larger availability of high-calorie food and the decreasing levels of physical activity results in an increasing spread of obesity and metabolic disorders in developed countries. One of the most important is prediabetes, which has been defined as a modern epidemic [153]. The consequences of prediabetes go beyond simple glycemic dysregulation. This multifactorial metabolic disorder can lead to many complications, like microvascular diseases such as neuropathy, nephropathy, and retinopathy, as well as macrovascular diseases like stroke, coronary artery disease and many others [154, 155]. The presence of prediabetes increases up to 10-fold the risk of developing type 2 diabetes (T2D), one of the most spread conditions worldwide [154]. Several studies have highlighted the potential of early detection and treatment of prediabetes in the prevention of these complications [153, 156]. The most reliable screening tool available is the *oral glucose tolerance test* (OGTT) which monitors the blood plasma glucose concentration after 2 hours from the intake of 75 g of glucose in water solution. The World Health Organization (WHO) distinguishes three types of prediabetes [157] based on the blood plasma concentration of glucose in fasting condition (*fasting plasma glucose*, FPG) and after two hours from the intake of the solution of glucose (*OGTT glucose*):

1. *Impaired fasting glucose* (IFG): FPG = 110 - 125 mg/dL (6.1 - 6.9 $mmol/L$)

2. *Impaired glucose tolerance* (IGT): OGTT glucose = 140 - 200 mg/dL (7.8 - 11.0 $mmol/L$)

3. IFG/IGT - both conditions together

The OGTT test, however, is time-consuming and difficult to reproduce [158]. Therefore, infrared spectroscopy of human blood serum or plasma can potentially offer a faster and reliable diagnosis. The potential of FT-IR and FRS spectroscopy is first investigated for the three prediabetes conditions in the KORA-FF4 cohort and their classification efficiency is then compared with the efficiency of common clinical parameters connected with prediabetes.

4.3.1 Comparison of FRS and FT-IR fingerprints of prediabetes in KORA-FF4

In the previous sections, FT-IR and FRS spectroscopy applied to human blood plasma has been shown to provide high classification efficiencies for individuals affected by diabetes. Prediabetes is an asymptomatic condition always antecedent to the development of type II diabetes. In this section, the potential of both techniques is explored for the three types of prediabetes listed above in the KORA-FF4 cohort via SVM binary classifications with all normal glucose tolerance (NGT) individuals (FPG < 110 mg/dL and OGTT glucose < 140 mg/dL) and with matched NGT controls. The binary classifications via the FT-IR data are first addressed in full-spectrum and then in the same spectral range covered by FRS (1000 - 1500 cm^{-1}) and compared with the classifications via the FRS data in the frequency and in the time domain.

The NGT individuals (> 1300) outnumber the cases (Figure 4.13a), which explains why the SVM coefficients found for the classifications with all NGT controls on the FT-IR in full-spectra are featureless (see section 4.2.3) and the AUCs are unexpectedly higher than the one found for diabetes (Figure 4.13b, c). The SVM classifications are performed also with NGT controls matched done for age, gender, CRP concentrations, hypertension and high blood lipids. After matching, the number of controls is the same as the number of cases and the AUCs found for the classification of IFG/IGT individuals are slightly smaller than for diabetes (Figure 4.13b). This is to be expected since the IFG/IGT is the latest stage of prediabetes and the closest to type II diabetes. Only a few cases are symptomatic for the IGT and IFG conditions which, if classified with matched NGT controls, return AUCs with average values of about 52% and 60% respectively. The SVM coefficients obtained in the classifications with matched NGT controls show that the features of IGT are spread in the whole spectral range, with larger signatures from proteins (1500 - 1750 cm^{-1}) and lipids (1800 cm^{-1}), while the main spectral signatures of IFG are associated with carbohydrates and proteins' glycosylations (1000 - 1250 cm^{-1}, Figure 4.13c). The SVM coefficient of IFG/IGT presents the features of both isolated conditions and is, therefore, better classified.

The same analysis is performed on the FT-IR spectra in the same spectral coverage as FRS (1000 - 1500 cm^{-1}) for the comparison of the two techniques (Figure 4.14). The SVM coefficient found in full-spectra for IGT shows that the main features responsible for its classification are the Amide bands of proteins and the signatures of lipids. Since the spectral signatures of IGT are not in the spectral range considered and contribute also to the signature of the IFG/IGT condition, it is not surprising that the average value of the AUCs found in the reduced spectral coverage for IGT and IGT/IFG are lower than in full-spectra, even though it is not statistically significant because of the large standard deviations.

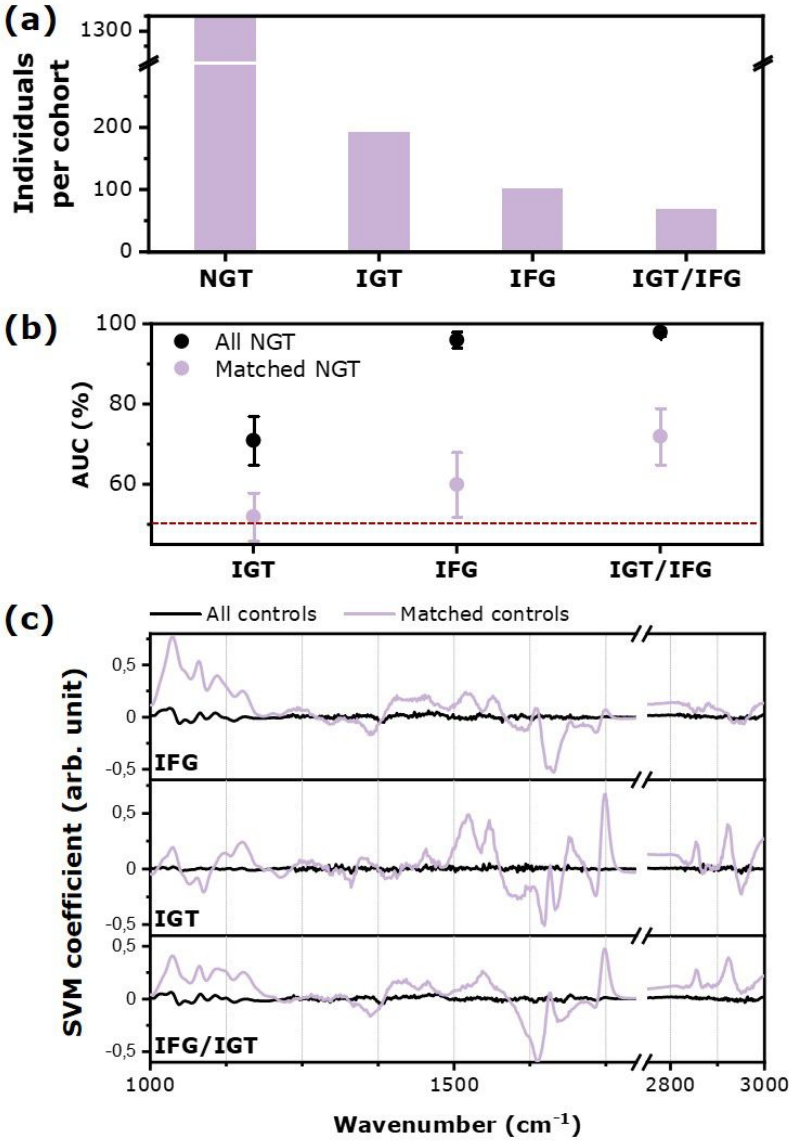

Figure 4.13: SVM binary classification of prediabetes conditions with matched and unmatched NGT individuals on the FT-IR full-spectra of KORA-FF4 cohort. (a) The number of individuals for each cohort highlights that the NGT individuals outnumber the prediabetes cases. (b) The classification efficiencies are unexpectedly higher than what was observed for diabetes (91 ± 3 %) for the classification of IFG and IGT/IFG conditions with all NGT individuals. The classification of all conditions with matched NGT controls returns increasing AUCs for IGT, IFG and IGT/IFG, which is expected considering that the blood plasma samples have been collected in a fasting conditions and that IGT is due to the glucose impairment after the intake of glucose, while IFG appears in a fasting condition. (c) The SVM coefficients corresponding to panel (b) are featureless in the classifications with all NGT individuals because of the large difference in the number of cases and controls (section 4.2.3). The SVM coefficients obtained from the classifications with matched NGT controls highlight that the main signatures of IFG are due to glucose and glycosylated proteins, while the spectral signatures connected with IGT are mostly due to proteins and lipids. The signature of IFG/IGT is a combination of the features of the isolated IFG and IGT conditions. Acronyms: SVM - support vector machine; AUC - area under the curve; NGT - normal glucose tolerance; IGT - impaired glucose tolerance; IFG - impaired fasting glucose.

The average AUCs found for the classifications with matched NGT controls on the FRS data, both in the frequency and the time domain, are higher than the values found for FT-IR in the same spectra coverage (Figure 4.15a). However, because of the small number of cases, the standard deviations are too large to have a significant difference. Further investigations on a larger number of cases are needed to address if FRS outperforms FT-IR spectroscopy in the detection of prediabetes. The higher AUCs found from the classifications of FRS data in the frequency domain compared to the FT-IR data agree with the higher number of features in the SVM coefficients derived from the FRS data in the frequency domain for the same type of prediabetes (Figure 4.15b).

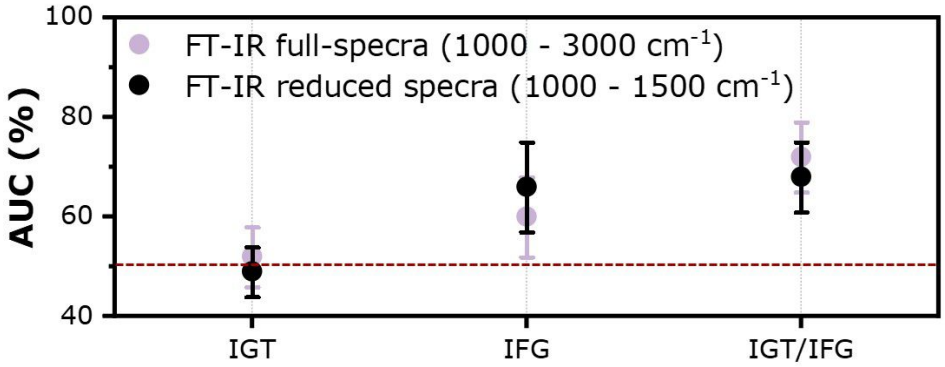

Figure 4.14: SVM binary classification of prediabetes conditions with matched and unmatched NGT individuals on the FT-IR full-spectra and reduced spectra of KORA-FF4 cohort. The AUCs found in the full-spectrum (1000 - 3000 cm^{-1}) and in the reduced spectra (1000 - 1500 cm^{-1}) are comparable within the standard deviations. However, the AUC of IGT and IGT/IFG are slightly smaller in the reduced spectral range which does not cover the main signature seen for IGT in Figure 4.13c. Acronyms: SVM - support vector machine; AUC - area under the curve; NGT - normal glucose tolerance; IGT - impaired glucose tolerance; IFG - impaired fasting glucose.

The SVM binary classifications of the FRS data in the time domain are performed as explained in section 2.3.2 and allows the identification of the optimal time window for each condition, namely where the AUC reaches the maximum value for that classification. As done above for FT-IR in full-spectra and the FRS analysis of endpoint medical conditions (section 4.2.2), the binary classifications of the three types of prediabetes are performed with all unmatched NGT individuals as well as with matched NGT controls. As for the endpoint diseases, the optimal time windows found in the classifications with unmatched controls are different compared to the optimal time windows obtained for the classifications with matched controls (Figure 4.16a).

Because of the small number of cases, the trends of the AUCs along the time axis are noisy and, for the classifications with matched controls, are affected by higher overfitting rates (Figure 4.16b). For example, the optimal time window identified from the algorithm for the classification of IGT with matched controls are affected by overfitting and a window between 1 and 4 ps, where the overfitting rate is lower, provides a more robust value for the AUC (around 60%). Similar conclusions as for the endpoint medical conditions investigated in section 4.2.2 can be drawn also for IGT, IFG and IFG/IGT, such as that the classifications with unmatched controls return the highest AUCs close to time zero (0 - 2 ps for IFG and IFG/IGT and at 0 - 4 ps

for IGT), while for the classifications with matched controls the AUCs reach the highest value for that condition in more confined temporal windows. In other words, the classifications with unmatched controls are due to the temporal signatures specific to the target medical condition, isolated via the classifications with matched controls, as well as to non-specific signatures arising around 0 *ps*.

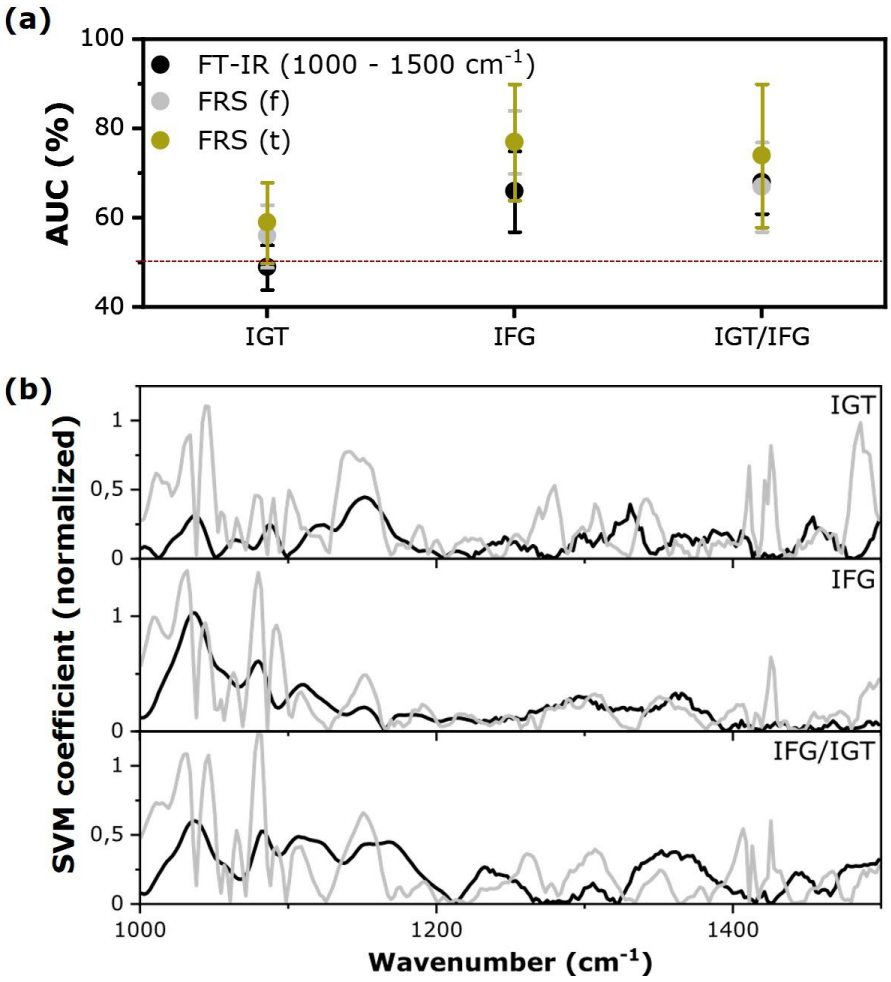

Figure 4.15: SVM binary classification of prediabetes conditions with matched and unmatched NGT individuals on the FT-IR and FRS data of KORA-FF4 cohort. (a) The average AUCs obtained via FRS are generally higher than for FT-IR in the reduced spectral coverage (1000 and 1500 cm^{-1}). However, the standard deviations are large because of the low number of cases. (b) The absolute values of the corresponding SVM coefficients found for FRS in the frequency domain show both common and extra features compared to the coefficients found via FT-IR in the same spectral coverage, in agreement with the higher AUC values. Acronyms: SVM - support vector machine; AUC - area under the curve; FRS - field-resolved spectroscopy; NGT - normal glucose tolerance; IGT - impaired glucose tolerance; IFG - impaired fasting glucose; FRS(f) - FRS signal in the frequency domain; FRS(t) - FRS signal in the time domain.

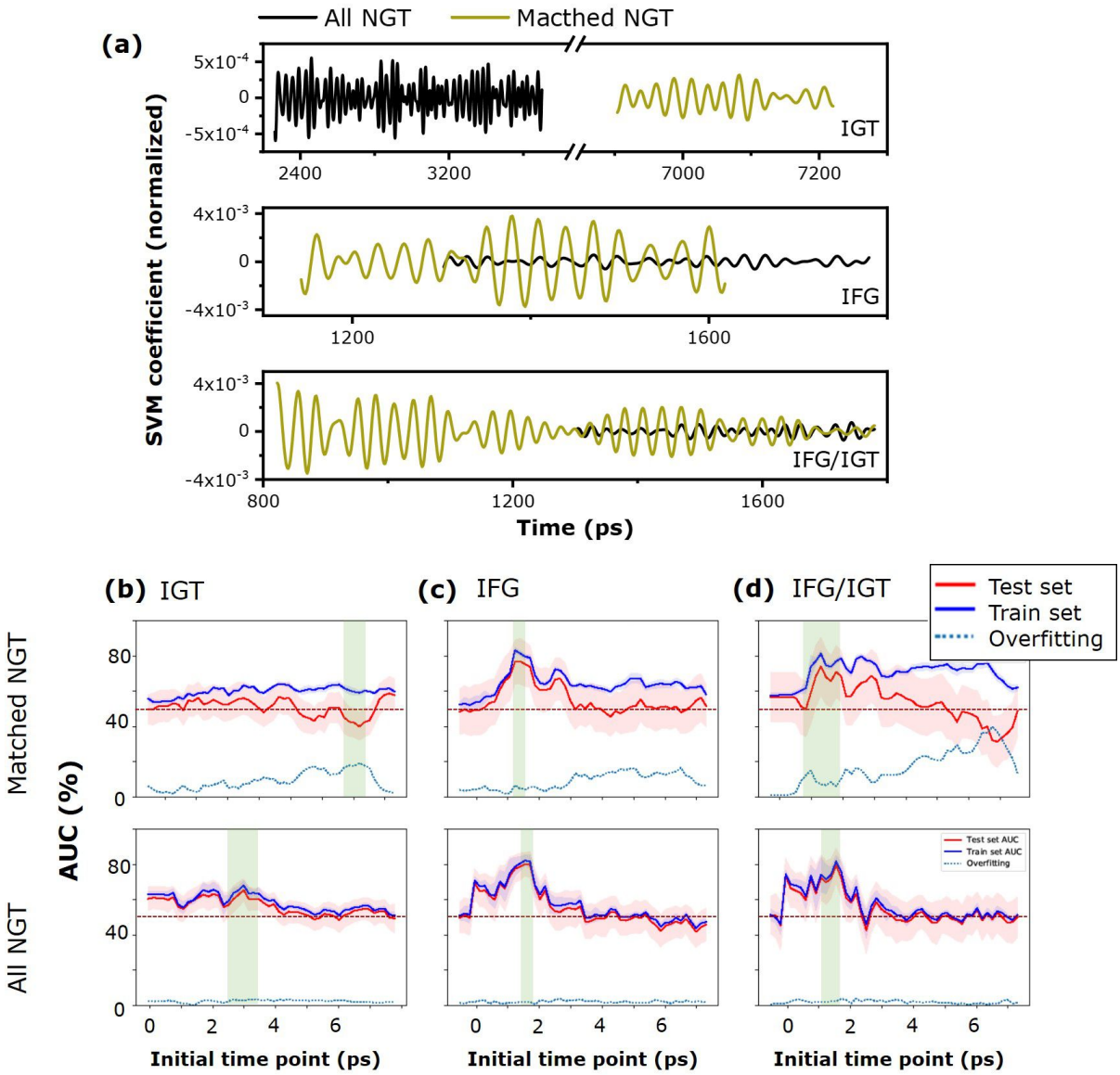

Figure 4.16: SVM coefficients of all prediabetes conditions with matched and unmatched NGT individuals of the EMFs of KORA-FF4 cohort. (a) The SVM coefficients are reported in the respective identified optimal time windows, which are different for the classifications with matched and unmatched controls for all prediabetes conditions. (b) The AUCs trends along the time axis are reported for diabetes and all prediabetes conditions for the classifications with matched (upper panels) and unmatched (bottom panels) NGT controls. The red dashed lines shows the AUCs= 50%. As for the endpoint medical conditions, the AUC can have values close to the maximum reached for that classification also outside the optimal time window (green shaded areas). In the classifications with unmatched controls, the AUCs have comparable values in the same optimal temporal windows found for the classifications with matched controls. However, the AUCs are always higher for unmatched controls around 0 *ps* due to the non-specific contributions of factors correlating with the target condition. Acronyms: SVM - support vector machine; AUC - area under the curve; FRS - field-resolved spectroscopy; EMF - electric-field resolved molecular fingerprint; NGT - normal glucose tolerance; IGT - impaired glucose tolerance; IFG - impaired fasting glucose.

The blood plasma samples of KORA-FF4 have been collected in fasting conditions, therefore, the IFG signatures, a condition that implies a glucose impairment during fasting, are stronger than for IGT, for which the impairment happens only after the intake of glucose. It s therefore expected that IFG is the leading condition also in the classification of IGT/IFG. The spectral signatures found via FT.IR and FRS data in the frequency domain are highly comparable with each other. In particular, the IR fingerprints show that the spectral signatures of IGT arise mostly from proteins and lipids, highlighting the need to identify the roots of prediabetes beyond glucose concentration as this can allow a timely detection even before the glucose impairment. The spectral signature at low frequencies identified for the IGT condition has a different shape, and probably a different molecular origin, compared to the same features in the SVM coefficient of diabetes and IFG. Unfortunately, the molecular attribution of these signatures is not straightforward as many biomolecules associated with IGT have spectral signatures that might correspond to the ones identified in this study. For example, the metabolomics study on the KORA-S4 cohort reported in [159] identifies different metabolites connected with IGT, among which well-known diabetes biomarkers (HbA1c, glucose and insulin) as well as metabolites with no previously known connection with diabetes and prediabetes, such as glycine (IR signatures between 1250 and 1750 cm^{-1}), lysophosphatidylcholine (18:2) (1000 - 1250 cm^{-1} and 1750 - 3000 cm^{-1}) and acetylcarnitine (1000 - 1250 cm^{-1}). The study also reports an AUC of 63% for the detection of IGT via the three metabolites listed, which is similar to the AUC obtained via FRS spectroscopy. The spectral signature of IFG is mostly due to glucose and glycosylated proteins with some contribution in the protein region, different than what was observed for IGT. However, the molecular interpretation is beyond the possibility of this analysis and further studies combined with other techniques providing molecular information are needed. In general, IR spectroscopy has the advantage of characterizing all molecules of the sample under investigation thus providing a tool to fast and reliably identify the pattern of changes in the concentration of any molecule connected with prediabetes.

To the best of our knowledge, only another published study uses infrared spectroscopy for the detection of prediabetes in human blood and is based on ATIR spectroscopy combined with machine learning algorithms not discussed in this dissertation [118]. The study reports high classification rates, however, it is based on a small sample set of only 50 cases and, more importantly, the three different prediabetes conditions are classified together, which is not optimal as the three are associated with different changes in the chemical composition of blood, as unveiled by the metablomics analysis reported in [159]. Moreover, in [118], the controls are not matched to the cases and the analysis can potentially be influenced by other factors correlating with prediabetes, as it is potentially the case also in the study reported here for which the AUCs found for the classification of prediabetes with unmatched controls are higher than the AUC found for diabetes. A fundamental difference with [118] is that the study reported here is based on the same classifier with the sole purpose of comparing the performances of FT-IR and FRS spectroscopy for the classification of different medical conditions, therefore no algorithm optimization has been performed for each particular medical condition. In general, the detection of prediabetes conditions via IR spectroscopy returns promising results, highlighting the potential of this approach. FRS outperforms FT-IR spectroscopy in the same spectral coverage, especially for the detection of IGT, not well classified via FT-IR, for which it returns an AUC of 60% with low overfitting rates. However, a more extensive study on a larger number of cases is necessary to establish the best performances of both techniques.

4.3.2 Classification of prediabetes via FRS and FT-IR fingerprints combined with clinical parameters

Despite promising, the performances of the FT-IR and FRS spectra of human blood plasma for the detection of the different types of prediabetes need to be compared with the efficiencies of other techniques and clinical parameters. Several easily measurable clinical parameters are indeed connected with diabetes and prediabetes conditions, such as glucose, glycohemoglobin (HbA1c) and insulin. In this section, the classification efficiency of the fasting plasma concentrations of these three biomolecules, obtained via the clinical analysis of the KORA-FF4 samples, are compared to the efficiencies of FT-IR and FRS data for the classification of the three prediabetes types. In particular, this is performed via merging the concentrations of glucose, HbA1c or insulin to the FT-IR and FRS data and see how they affect the classification efficiency: if the classifications return similar AUCs as the IR data alone, it can be concluded that the clinical parameters do not add any information about the prediabetic status to the IR fingerprints; if the AUCs are higher than the IR fingerprints alone, it means that the clinical parameters are better classifiers than the spectroscopic data.

Figure 4.17a shows that, if the fasting insulin concentrations are merged to FT-IR full-spectra, the AUCs of all conditions drop to 50% with large standard deviations. Therefore, insulin is a confounding factor. If the FT-IR data are merged with the fasting HbA1c concentrations, the efficiencies increase up to 80% and 95% for IGT and IFG(/IGT) respectively. When both insulin and HbA1c are merged with the FT-IR spectra the AUCs have very large standard deviations, meaning that including insulin is again detrimental for the classification. Merging the fasting glucose concentrations with the FT-IR data returns AUCs up to 99% for prediabetes conditions involving IFG. This is expected since the fasting glucose concentrations are the parameters used to define IFG cases. Therefore, any classification of IFG and IFG/IGT including this clinical parameter will necessarily result in the highest classification efficiency. This also explains the high AUCs found including fasting HbA1c since it correlates with fasting glucose concentrations (Figure 4.17a). On the other hand, IGT is defined from the OGTT test based on the glucose concentration after the intake of this molecule in solution (not in fasting conditions). Therefore, both fasting glucose and HbA1c can be used as classifiers for IGT and are better predictors than the FT-IR spectra.

The same conclusions can be achieved via the FT-IR reduced spectra (1000 - 1500 cm^{-1}, Figure 4.17b). FRS in the frequency domain delivers similar AUCs as FT-IR in the same spectral coverage, with the main difference for the classification of IFG, for which FRS performs slightly better than FT-IR giving similar AUCs without clinical parameters as including insulin alone or with HbA1c. However, merging the FRS data with these clinical parameters does not return higher AUCs than the FRS data alone, therefore it can be concluded that insulin and HbA1c do not add any information about IFG to the FRS data. In the time domain, the EMFs traces merged with the clinical parameters return AUCs with very large standard deviations, except for IGT: if the EMFs are merged with insulin concentration the AUC is unexpectedly high, even higher than what has been reported in the literature for the detection of diabetes via insulin alone [160]; if the EMFs are merged with glucose the AUC is as high as for the classification via the EMFs alone.

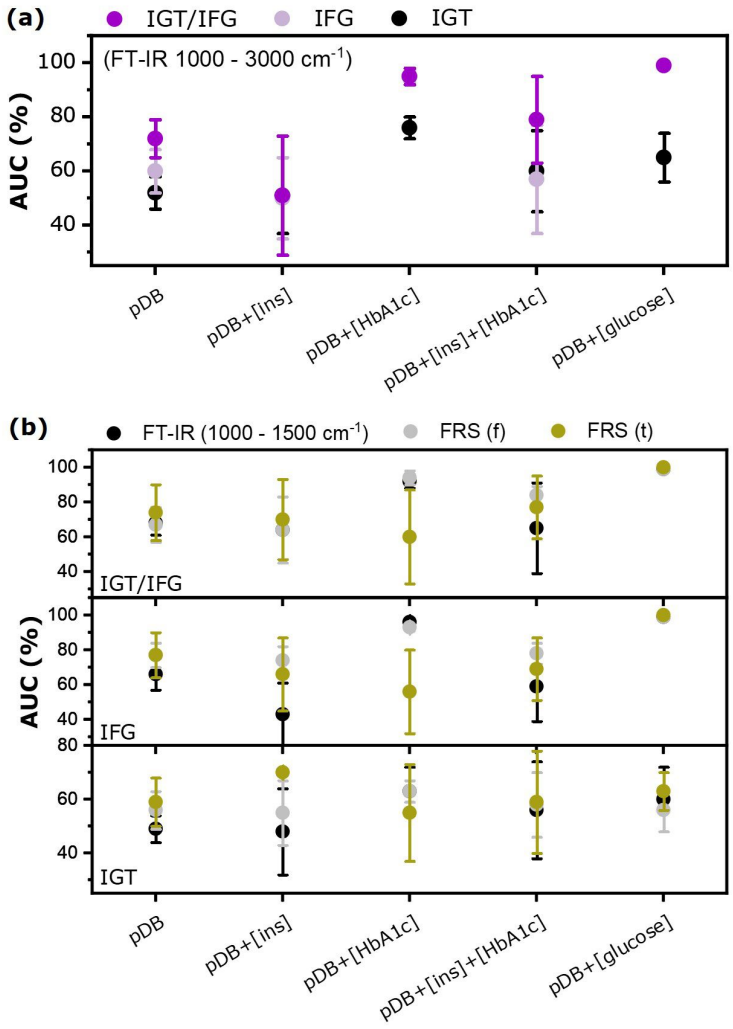

Figure 4.17: SVM binary classification of prediabetes conditions with matched NGT individuals on the FT-IR and FRS data merged with clinical parameters of KORA-FF4 cohort. (a) The AUCs are reported for the FT-IR data in full-spectra and (b) for the FT-IR data in the reduced spectral coverage (1000 - 1500 cm^{-1}) and for the FRS data, both in the time and frequency domain (pDB in the legend; clinical parameters: [ins], [HbA1c] and [glucose]). The lower AUCs found with insulin highlights that it is not a good predictor. The AUCs are close to 99% for the classification of the IR data merged with glucose and HbA1c if IFG is present as expected since the fasting glucose concentration is the parameter used to identify IFG and HbA1c correlates with it. IGT is identified via OGTT, therefore the fasting glucose concentration can be used as a predictor and it performs better than the FT-IR spectra. The same conclusions can be drawn for FT-IR and for the FRS data, except for the AUCs found from the EMFs traces for the detection of IGT: if the EMFs are merged with insulin concentration, the AUC is unexpectedly high (higher than what has been reported in the literature for the detection of diabetes via insulin alone [160]); if the EMFs are merged with glucose, the AUC is as high as for the classification via the EMFs alone highlighting that FRS has at least the same predicting power as fasting glucose. Acronyms: SVM - support vector machine; AUC - area under the curve; FRS - field-resolved spectroscopy; pDT - prediabetes; NGT - normal glucose tolerance; IGT - impaired glucose tolerance; IFG - impaired fasting glucose; FRS(f) - FRS signal in the frequency domain; FRS(t) - FRS signal in the time domain; [ins] - fasting plasma concentrations of insulin; [HbA1c] - fasting plasma concentrations of glycohemoglobin; [glucose] - fasting plasma concentrations of glucose; OGTT - oral glucose tolerance test.

To the best of our knowledge, this is the first analysis reporting the efficiency of prediabetes detection via IR spectra merged with clinical parameters. From the results reported, it can be concluded that insulin is a confounding factor and merging its concentrations with the IR data result in worse or unexpected classification outcomes. Fasting plasma glucose and HbA1c concentrations are better classifiers than FT-IR for IGT, which is not surprising considering the low AUC obtained via FT-IR. However, the binary classifications via FRS data, both in the time and frequency domain, return higher classification efficiencies for IGT with similar AUCs with and without merging the glucose concentrations to the data. This highlights that FRS is a better classifier than FT-IR and that it is at least as good as fasting plasma glucose concentration alone.

The classification efficiency of plasma glucose, insulin and HbA1c have been reported in the literature. In particular, the AUCs obtained from fasting capillary glycemia have been identified from a large cohort (> 4000 individuals) as 91 ± 1% and 75 ± 1% for the diagnosis of diabetes and IGT respectively [161], comparable with what has been reported from the classifications via FPG ad HbA1c (AUCs of 90% for diabetes and 65-70% for IGT) [162]. Compared to these studies, the AUC found here for the classification of the 288 diabetes cases with all controls is similar (91 ± 3% for FT-IR and 88 ± 4% for FRS), while for the classification of the 200 IGT cases the AUC found via FRS is about 60%, but it is expected to reach higher values from the analysis of a larger number of cases (section 4.2.3). The study [163] shows that fasting hyperinsulinemia itself has a primary role in the pathogenesis of diabetes, but our study shows that fasting insulin concentration is not a good predictor. The insulin sensitivity index at 0 and 120 minutes during the OGTT test have been reported to return an AUC of 78.5% for the prediction of diabetes [160], which is lower than what is found via IR spectroscopy, consistently with the observation that insulin does not provide additional information to the spectra.

Different techniques not based on IR spectroscopy have been applied to increase the detection efficiency of prediabetes compared to the tests currently available. Mass spectrometry is one of these techniques and it has been reported to reach an AUCs of 84% for the detection of IGT, considering 99 metabolites and 22 lipoproteins associated with the 2-hour glucose or insulin concentrations [164], and of 87% for the detection of IFG from 23 metabolites, mostly amino acids, carnitines and phospholipids [165]. Compared to mass spectrometry, IR spectroscopy has the advantage of addressing all molecular constituents in one fast measurement. The AUCs found via FRS spectroscopy are comparable for the detection of IFG, but lower for IGT thus stressing the importance of addressing the best performances of this technique via studies on larger cohorts (section 4.2.3) as well as to improve its performances via the planned technical developments which will allow FRS to detect a larger dynamic range of molecular concentrations and will broaden the spectral coverage so to cover an even larger spectral coverage than the one of FT-IR.

One of the main limitations of these studies is that there is no way to access the actual health status of an individual which is addressed via the current golden standard method. In particular, all the studies listed above, including the one reported here, are based on the OGTT outcomes. Despite OGTT is the golden standard for the detection of diabetes and prediabetes, it has been reported to have a low reproducibility because of the influence of sampling timing, diet, exercise, age, gastrointestinal factors and stress before the test [166]. The study [167] reports the AUCs of FPG and OGTT taking as a baseline a clinical model based on age, gender, ethnicity, BMI, blood pressure, FPG, HDL and family history of diabetes giving an AUC between

80-85% for OGTT. Therefore, it is really difficult to assess if the miss-classifications are due to a wrong assessment of the test used or to the actual true status of the individual miss-classified via the OGTT test. Longitudinal studies of the same individuals over time might help to address the actual predicting power of the OGTT test compared to different techniques. To this purpose, the KORA cohort, which has been monitored at different time points, offers again a unique advantage compared to other studies. The FT-IR and FRS analysis of the KORA-F4, the same individuals as KORA-FF4 sampled about seven years before, will be performed in the near future and compared with the presented data for the assessment of the efficiency of IR spectroscopy in a longitudinal fashion.

4.4 Concluding remarks

This chapter addresses the molecular changes induced by several common medical conditions in human blood biofluids via vibrational spectroscopy. Both FT-IR and FRS spectroscopy are used to record a snapshot of all molecular constituents simultaneously of each sample to compare the chemical composition of the human blood biofluids of individuals symptomatic to a target disease with the one of non-symptomatic individuals. Most of the analyses of vibrational spectroscopy for the detection of diseases reported in the literature are on clinic-based case-control cohorts, while in this chapter both FT-IR and FRS spectroscopy are applied on a cross-sectional population-based cohort, KORA-FF4, which provides the advantage to allow identifying the IR signatures of any condition or phenotype in the general population represented.

In the first section, unsupervised algorithms and PCA have been applied on the whole KORA-FF4 cohort to stress the impact of inflammation and age on the between-person spectral variability of FT-IR spectra already observed in the previous chapter for healthy individuals, which is expected because of the high similarity between the loading vectors of healthy individuals and the whole KORA-FF4 cohort. The origin of the between-person variability of the general population is different for the spectral range covered by FRS, between 1000 ad 1500 cm^{-1}, which is mainly affected by gender and inflammation, as seen for the healthy individuals, as well as from diabetes.

Despite the large effect of common parameters on the between-person spectral variability, it is possible to isolate the features of medical conditions via binary classifications. In particular, SVM is applied to identify the IR fingerprints and diagnostic power of FT-IR and FRS spectroscopy on human blood biofluids for each endpoint and intermediate medical condition known for KORA-FF4 and in the independent L4L cohorts. The comparison of independent cohorts highlights the importance of matching cases and controls to isolate the disease-specific signatures and classification efficiencies for each medical condition, otherwise affected by the correlating parameters and comorbidities. For example, diabetes and hypertension correlate with each other and are shown to mutually influence their SVM coefficients if classified with unmatched controls. The importance of matching to address if a given condition influences the IR fingerprints of the biosample under investigation is highlighted by the binary classification of asthma with unmatched controls, for which the AUC is higher than 50%, and with matched controls, which returns no separation showing that the separation seen with unmatched controls arises from factors correlated with this medical condition and it is not specific to asthma. This study stresses the importance of matching cases and controls to isolate disease-specific IR

signatures, but the analysis of larger cohorts and more medical conditions are fundamental to give a general validity to these conclusions and to define a standard procedure for matching cases and controls for each disease.

A drawback of matching is the reduced number of controls. The impact of the number of cases and controls on the classification efficiency has been shown via learning curves reporting the AUCs for increasing number of cases for the FT-IR spectra of diabetes and hypertension in the KORA-FF4 cohort highlighting how the classifications with matched controls return systematically lower classification efficiencies than with more numerous unmatched controls because of the combined effect of reducing the unspecific contributions and of reducing the number of controls. It has been shown that there is a minimum number of cases required to reach robust classification efficiencies and IR signatures which depends on how strongly a given medical condition impacts the IR spectra, namely on the corresponding AUC. In particular, the study reported here shows that the minimum number of cases required to address the efficiency of SVM binary classifications of FT-IR human blood plasma fingerprints is higher than 380 for AUCs lower than 65%, as for hypertension, and about 130 for medical conditions giving high AUCs, as for diabetes, which is comparable to the conclusions reported in the literature for another classification algorithm for the Raman spectra of single cells [150] which identifies 75-100 cases as the minimum number to achieve good but not perfect classifications. These conclusions might have a general validity for similar studies to address the reliability of AUCs of 50%.

A high number of cases guarantees high statistical power to the study [144], namely robust classifications and IR fingerprints. It is therefore expected that the conditions with the most numerous cases return also higher AUCs, which is the case for diabetes, hypertension and high blood lipids. Stroke and heart attack cohorts count a small number of cases, but the classification efficiencies are higher than 50% thus making the analysis of larger cohorts of cases important to address the optimal classification efficiencies of these conditions and the robust identification of their IR signature which could potentially help to timely identify people at high risk. The AUCs found for COPD, former cancer and asthma cohorts are close to 50% and the comparison with the classification efficiency for the same number of hypertension cases shows that for these three conditions a very large number of cases is needed to reach robust classifications and are expected to return low AUCs. Therefore, IR fingerprinting of human blood plasma might not be suited for their classification.

The asymptomatic intermediate condition of prediabetes is also investigated in this study via both FT-IR and FRS spectroscopy on the KORA-FF4 cohort. In particular, three types of prediabetes have been investigated: IFG, glucose impairment during fasting; IGT, glucose impairment after the intake of a glucose solution; IFG/IGT, glucose impairment both in fasting and not fasting conditions. Since the blood plasma samples of KORA-FF4 have been collected in fasting conditions, the IFG signatures are stronger than for IGT and it is the most important signature in the IFG/IGT condition.

The average AUCs identified for the classification of prediabetes via FRS are higher compared to the ones found for FT-IR in the same spectral range, showing that FRS performs better for the detection of these conditions. Moreover, by increasing the number of cases, higher AUCs can be obtained for the optimal performances of FRS spectroscopy. The AUCs of the endpoint medical conditions investigated here are comparable between FT-IR and FRS in the same spectral coverage, which is a promising result for the current version of FRS affected by

the day-to-day dependence discussed in the previous chapters. A major advantage of the time resolution of FRS over frequency-resolved techniques is that it allows the identification of the disease-specific signatures, visible in the classifications with matched controls, in narrower and separate temporal windows compared to the non-specific signatures due to factors that correlate with the target condition.

Most of the medical conditions are affected by lifestyle and diet [137–139] and IR spectroscopy can be a powerful method for their timely and simultaneous diagnosis, as well as monitoring, and could potentially help to identify individuals at high risk which can modify their lifestyle preventing or delaying the medical condition. A molecular interpretation of the IR fingerprints of medical conditions is also attempted in this study, but it is beyond the purpose of this dissertation as it would require the comparison with the analysis of the same samples via techniques able to address the molecular changes, similarly to what is reported in the next chapter for lung cancer for which the FT-IR spectra are compared with the mass spectrometry analysis of the same samples. In particular, the spectral signature of diabetes is stronger at low frequencies (1000 – 1250 cm^{-1}) where the main contributions come from carbohydrates and proteins' glycosylations, in agreement with the well-known consequences of the glycemic dysregulation on these biomolecules [129, 130]. The signatures of hypertension are spread in the whole spectrum. Urea and creatinine have been reported as potential biomarkers for hypertension in [136] and have both spectral features between 1000 and 1800 cm^{-1}, but further analyses are needed to establish the molecular origin of the fingerprint of hypertension for which this is, to the best of our knowledge, the first study addressing the corresponding FT-IR signature in human blood serum and plasma. The main features responsible for the classification of individuals with high blood lipids arise from proteins and lipids and are probably due to lipoproteins [134], while the smaller contributions at lower frequencies could be due to triglycerides and glycerol [140]. The strong signatures at higher frequencies found for heart attack are mostly due to lipids (1750 – 3000 cm^{-1}), in agreement with the increased risk of heart attack for higher blood concentrations of low-density lipoprotein (LDL) [132, 133] and with the comparison of the FT-IR spectra of HDL and LDL with the spectrum of human blood serum in [134]. The observed lower fasting blood glucose of individuals with episodes of heart attack [135] could explain the signature at low frequency. The IR signatures of stroke are also spread on the whole spectral range, with the main contributions from proteins and lipids, probably originating from lipoproteins [141], while the features at low frequency could be due to amino acids [142] and organic and inorganic metal complexes [143] associated with a high risk of stroke. Many studies reported in the literature want to address the best biofluid between human blood plasma for the detection of diseases [114, 127]. The study reported here on human blood plasma (KORA-FF4 cohort) and serum (L4L cohort) via FT-IR spectroscopy does not identify evident advantages of one biofluid over the other for the detection of diabetes, heart disease, hypertension and asthma, which can be due to the similar chemical composition of the two blood biofluids [128].

All prediabetes conditions return similar signatures in the frequency domain for FT-IR and FRS fingerprinting, but the molecular attribution of these signatures is not straightforward as many biomolecules associated with IGT have spectral signatures that might correspond to the ones identified in this study. The metabolomics analysis of the KORA-S4 cohort reported in [159] identifies glycine (IR signatures between 1250 and 1750 cm^{-1}), lysophosphatidylcholine (18:2) (1000 - 1250 cm^{-1} and 1750 - 3000 cm^{-1}) and acetylcarnitine (1000 - 1250 cm^{-1}) as potential

biomarker for the detection of IGT which provide an AUC close to the AUC obtained via FRS spectroscopy. The spectral signature of IFG can be associated mainly with glucose and glycosylated proteins, as it can be expected from the literature [168], but amino acids and phospholipids might also contribute [165].

IR spectroscopy has been shown particularly promising for the detection of diabetes and prediabetes conditions. MIR spectroscopy has been previously proposed as new methods for the detection and monitoring of glucose concentration using the spectral signatures at 1082 and 1093 cm^{-1} [169]. The review [170] summarizes recent achievements in glucose monitoring via IR spectroscopy of human tissues using QCL sources. However, [171] underlines the hazard of addressing prediabetes only based on glycemia because it has a slow impact on the cardiovascular risks due to diabetes [172] and addresses the importance of monitoring other factors able to detect the asymptomatic condition of prediabetes. The IR fingerprints reported in this study show that the spectral signatures of IGT arise mostly from proteins and lipids, highlighting the need to identify the roots of prediabetes beyond glucose concentration as this can allow a timely detection even before the glucose impairment.

Similar studies based on vibrational spectroscopy for the detection of diabetes and prediabetes in human blood biofluids have been reported in the literature. For example, ATR-FTIR analysis of human whole blood combined with XGBoost algorithms has been applied on a very small number of cases (50 individuals) for the detection of diabetes, reporting classification efficiencies close to what was observed in this study [131]. The same technique has been applied for the detection of prediabetes (50 individuals) reporting higher classification rates than what has been seen in this study [118]. However, in [118] the three different prediabetes conditions are classified together, which is fundamentally wrong since different types of prediabetes lead to different changes in the chemical composition of blood [159], and the controls are not matched to the cases so the analysis can potentially be influenced by other factors correlating with prediabetes. A fundamental difference with [118] is that the study reported here aims at comparing the efficiencies of FT-IR and FRS spectroscopy and no algorithm optimization has been performed for any medical condition.

To the best of our knowledge, the analysis of prediabetes via IR spectra merged with clinical parameters is reported here for the first time highlighting that insulin is a worse classifier than IR spectroscopy, in agreement with the AUC of 78.5% for the prediction of diabetes reported in [160], while fasting plasma glucose, and the correlated HbA1c concentrations, are better classifiers than FT-IR for the detection of IGT. However, FRS performs at least as fasting plasma glucose concentration alone. In particular, the AUCs found here for the classification of the 288 diabetes cases with all controls is similar (91 ± 3% for FT-IR and 88 ± 4% for FRS) as compared to the values reported for fasting capillary glycemia (FCG) [161], FPG and HbA1c [162]. For the classification of the 200 IGT cases, the AUC found via FRS is lower than the one ones reported for FCG [161], FPG and HbA1c [162], but it is expected to reach higher values from the analysis of a larger number of cases.

Different techniques have been applied to similar studies, such as mass spectrometry, which has been reported to return an AUCs of 84% for the detection of IGT [164] and of 87% for the detection of IFG [165]. The AUCs found via FRS spectroscopy in this study are comparable for the detection of IFG, but lower for IGT, stressing again the importance of addressing the best performances of this technique via studies on larger cohorts and to further boost the classification efficiency of the FRS technique via detecting a larger dynamic range of molecular

concentrations.

All the studies about diabetes ad prediabetes, including the one reported here, are based on the OGTT outcomes, which have been reported to have a low reproducibility [166] and to have an AUC of 80-85% compared to an independent clinical model [167]. It is therefore impossible to assess if the miss-classifications are due to a wrong assessment of the technique used or to the actual true status of the individual miss-classified via the OGTT test. Longitudinal studies like the KORA cohort might help to address the actual predicting power of different techniques. The FT-IR and FRS analysis will be performed in the near future on the samples collected from the same individuals at a different time point to assess the efficiency of IR spectroscopy longitudinally.

Ultimately, it can be concluded that, despite the large between-person spectral deviation due to age and inflammation, IR spectroscopy can detect the effect of diseases on human blood biofluids, particularly for diabetes, prediabetes, hypertension and high blood lipids. Larger cohorts are necessary to address the best performances for the detection of prediabetes, heart attack and stroke, while IR spectroscopy is probably not suited for the detection of the signatures of asthma, COPD and former cancer patients. The classification efficiencies found for IR spectroscopy are comparable to the classifications reported for other techniques of vibrational spectroscopy as well as for other methods and are higher for FRS spectroscopy for all prediabetes conditions. Moreover, the time resolution of the FRS traces allows identifying time windows where the signal is disease-specific without the need for matching cases and controls which reduces the number of controls and, therefore, the statistical power of the study. Further developments aiming at improving the technical stability of the FRS setup, broaden the spectral coverage and increase the dynamic range of concentration aim at further improving the classification efficiencies of FRS spectroscopy.

Chapter 5

Chemical fractionation and infrared fingerprinting for cancer detection

This dissertation shows the potential of infrared spectroscopy for the detection of medical conditions in different settings. Infrared spectroscopy is a well-known technique increasingly applied for this purpose, but the molecular origin of the fingerprints of diseases mostly remains unexplored. This is one of the main reasons for the slow improvements in the field and the low acceptance of the method in clinics.

One of the main advantages of FT-IR spectroscopy is the direct access to all molecules in human blood plasma/serum. The large dynamic range of molecular concentrations in blood derivatives is an obstacle for the detection of minimal changes in the concentration of highly abundant as well as of low abundant molecules. Proteins make up about 80% of all molecules of human serum (slightly more in plasma), leaving about 11% of metabolites and 9% of inorganic salts [173]. Of all the proteins, human serum albumin (HSA) constitutes alone about 50-60% of the total concentration [174]. This large protein, mainly featuring alpha-helix secondary structure, strongly contributes to the amide bands of all fingerprints shown in chapters 3 and 4 and covers the signatures of less abundant proteins.

In this section, the potential of separating human serum and plasma in albumin-depleted and albumin-enriched fractions is explored for a deeper understanding of the infrared fingerprints in pilot studies based on smaller cohorts than what has been considered in the previous chapters. In particular, the signatures of three cancer types are investigated: lung (LCa), breast (BCa) and prostate cancer (PCa). The fractionation of human blood biofluids is expected to provide a better understanding of the spectral signatures of less abundant molecules covered by the features of albumin. The molecular understanding of these fingerprints is better addressed via comparing the FT-IR spectra with the analysis of the same samples via mass-spectrometry. Before going into details about the analysis of cancer, the protocol used for the fractionation of the human blood biofluids samples is presented and the reproducibility is discussed.

5.1 Fractionation protocol: steps and reproducibility

The large dynamic range of molecular concentration of complex biosamples such as human blood plasma and serum is a drawback in the analysis via FT-IR spectroscopy as highly abundant molecules have predominant signals that cover the signatures of less abundant molecules. A solution to this problem is to reduce the chemical complexity of the bio-sample by splitting it into different fractions. In particular, for human blood biofluids, the ideal solution would be to separate metabolites from proteins and further separate the protein in albumin-enriched and albumin-depleted protein fractions using a fractionation protocol compatible with IR spectroscopy.

Commercial fractionation kits and most of the chemical approaches introduce contaminants in relatively high concentrations and with low reproducibility [175, 176] or break molecular bonds [177]. The cheapest commercial kits for depletion or fractionation are based on filters, membranes and columns which often contain or require the use of chemicals that contaminate the samples [175], very inconvenient for spectroscopic applications, and are not fully optimized for samples with such a high chemical complexity as human blood plasma and serum. Highly sophisticated, reproducible and targeted antibodies-based kits are extremely expensive (> 1000 euros/sample), making their use even on small pilot studies costly [178]. Chemical fractionation by precipitation via alcoholic solutions is the method that suits the most spectroscopic applications. The approach used in this chapter is an adaptation of Cohn´s method based on the studies reported in [176, 179], which was designed during world war II under the urgent need of plasma-based biological medicines like albumin and IgG [176, 180, 181], extremely important for the treatment of different diseases [182, 183]. Cohn´s method is a widely-used protocol in the industry [184].

The developed protocol is performed at 4°C, always using the same number of samples to ensure reproducibility, especially for the concentration stage (Figure 5.1a). Adding 5% of NaCl 1.1 M solution and about 45% of high-grade ethanol induces the precipitation of most proteins and lipids, leaving in solution only albumin, very few other proteins and most of the metabolites apart for lipids. The albumin-depleted pellet is redissolved in water by vortexing it for 90 minutes. The remaining proteins are precipitated by adding 55% of high-grade methanol to the supernatant. The albumin-enriched pellet is redissolved in water by leaving it at 4°C for 60 minutes. All three fractions obtained are then placed in a vacuum concentrator for 3 hours to reduce the dilution and effectively remove the alcoholic components. The protocol just described provides more than the simple albumin-depleted protein fraction, allowing the isolation of a metabolite fraction without introducing unwanted molecules. However, the albumin-depleted protein fraction is not entirely redissolved in water; the residual pellet has been stored at -80°C for future analysis.

The technical reproducibility of the developed precipitation protocol has been tested via FT-IR spectroscopy on 8 serum quality controls replica (QC, intra-sample) and the performance on real human plasma samples has been tested on 40 individuals (between-person), 20 lung cancer patients and 20 cancer-free controls matched for average age, gender and smoking status. The FT-IR spectra of each fraction in lung cancer patients (Figure 5.1b) shows that the albumin-enriched fraction has the highest intensity among the fractions with a similar spectrum as for the full human plasma, which is expected because of the higher concentration of albumin. The spectra of the two protein fractions are similar, which stresses the importance of separating

the high abundant albumin from the other proteins to better resolve the two signatures. The metabolite fraction has a different average spectrum with comparable intensity to the one of the albumin-depleted fraction. The sum of the average absorption spectra of the three fractions reconstructs the average spectrum of full plasma and highlights that the usual attribution of the spectral signatures to different biomolecules (Table 2.1) is an over-simplification since the contributions from proteins are relevant also in the spectral region attributed to metabolites, and vice versa [62].

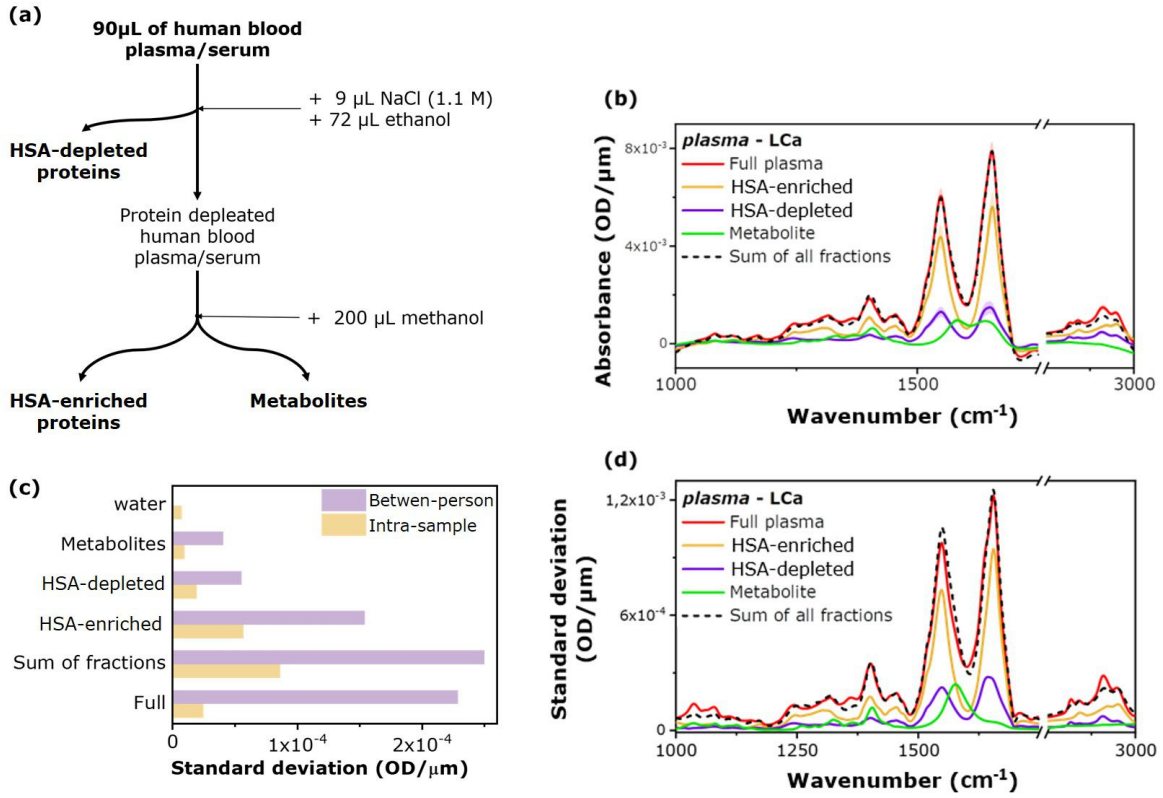

Figure 5.1: Outline and characterization via FT-IR spectroscopy of the proposed fractionation protocol. (a) The schematic outline of the fractionation protocol is depicted: 5% of NaCl 1.1 M and 45% of high-grade ethanol induce the precipitation of most proteins and lipids (HSA-deleted protein fraction), redissolved in water by vortexing it for 90 minutes; 55% of high-grade methanol leads to the precipitation of the remaining proteins (HSA-enriched pellet), also redissolved in water, leaving mostly metabolites in solution (metabolite fraction); the three fractions are placed in a vacuum concentrator for 3 hours to reduce the dilution and effectively remove the alcoholic components. (b) The FT-IR spectra of full human blood plasma, of the corresponding fractions and the sum of the spectra of each fraction are shown for 20 LCa patients. The intensity is higher for the HSA-enriched fraction, according to the higher concentration of HSA, which has a similar spectrum as for the HSA-depleted fraction. The metabolite fraction has a different average spectrum of comparable intensity to the one of the HSA-depleted fraction. (c) The corresponding total standard deviation of the spectra is reported for the 20 LCa cases (between-person) and 8 replicas of a serum QC (intra-sample) show that the most abundant molecules address the highest variability in the samples. (d) The between-person spectral variability is shown also resolved in the frequencies. Acronyms: HSA - albumin; LCa - non-small cell lung cancer; QC - quality control.

The standard deviation recorded for the FT-IR spectrum of each fraction is calculated and scaled according to the respective dilution (Figure 5.1c, d). The sum of the spectral standard deviation of each fraction obtained from QC replicas is higher than the spectral standard deviation of the full serum samples (Figure 5.1c, yellow bars), showing that the protocol increases the measurement error as it is expected since any sample manipulation inevitably increases the technical noise. However, the noise introduced by the fractionation protocol has a low impact on the standard deviation of the 20 lung cancer cases, for which the main source of noise is the between-person spectral variability (Figure 5.1c, d). The largest between-person spectral variability among the 20 lung cancer cases comes from the albumin-enriched fraction, in agreement with the high variability of the protein signatures observed for the KORA-FF4 individuals (section 3.2.2 and 4.1) and with the influence of age and inflammation on these signatures, both connected with a down-regulation of albumin [185, 186]. The sum of these standard deviations reproduces the variability of full plasma samples proving that the additional noise introduced with the fractionation is negligible compared to the variability due to the biological differences between individuals. The higher source of variability in the metabolite fraction is due to the vibrational signatures of EDTA, the anti-coagulant used for the withdrawal of the blood plasma samples, which will be shown to be informative for the detection of cancer. The high variability associated with EDTA can be expected based on of the multiple factors that can influence its concentration in the blood plasma, such as the variability in the tubes' coatings and of the volume of blood in the tube.

The molecular composition of each fraction can be tackled only by combining the spectroscopic fingerprints with a technique able to provide molecular-specific information [187]. A variety of *omics* techniques have recently emerged for in-depth investigation of several biofluids [188–192]. In particular, high-throughput mass-spectrometry proteomics for the analysis of human blood plasma [193] has been used to analyze the same plasma samples here investigated via FT-IR spectroscopy and it has been adapted and applied also to the analysis of the corresponding albumin-depleted fractions. Figure 5.2a shows the average mass-spectrometry intensities for the first most abundant proteins for the full plasma and the albumin-depleted protein fraction of the 20 lung cancer patients and of the 20 controls together. From the mass-spectrometry intensity, it can be calculated that the concentration of albumin in the albumin-depleted protein fraction is reduced to about 15% (Figure 5.2b), both in the real samples ad in the 8 QC replicas, meaning that the proposed fractionation protocol has an albumin-depletion efficiency of 85%. Mass-spectrometry shows that 99% of the dry mass of the albumin-enriched protein fraction count mainly 3 proteins: albumin (ALB), haptoglobin (HP) and alpha-1-acid glycoprotein (ORM1) [62]. Mass-spectrometry also shows that the applied fractionation protocol adds only marginal noise to the instrumental one, as described in [62] and observed via the FT-IR between-person spectral variability.

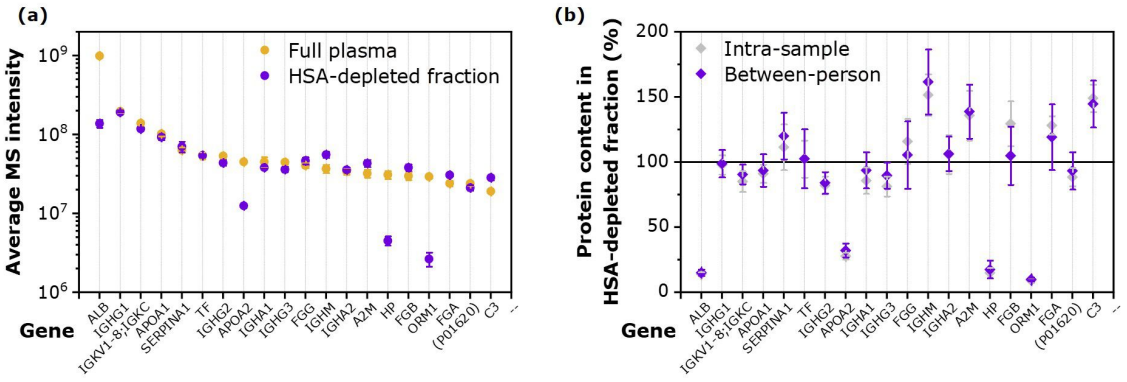

Figure 5.2: Mass-spectrometry proteins intensities of the full plasma samples of 40 individuals, the 8 QC full serum replicas and the respective HSA-depleted protein fractions. (a) The average MS intensities are reported for the most abundant proteins of full plasma and HSA-depleted protein fraction for the 20 LCa cases and the 20 controls in absolute average values. (b) The mass-spectrometry intensities of the most abundant proteins found in the HSA-depleted protein fraction are reported in percentage with respect to the intensities observed in the full plasma samples and are compared with the same values found for the 8 QC serum replicas. The proteins are labeled using the respective gene. The MS intensity of HSA (gene: ALB) found in the HSA-depleted protein fraction is 15% compared to the MS intensity found in the full serum and plasma samples showing that the fractionation protocol applied provides an HSA-depletion efficiency of about 85%. Among the most abundant proteins, three are depleted with the applied fractionation protocol, in particular HSA (ALB), haptoglobin (HP) and alpha-1-acid glycoprotein (ORM1), which make up the main constituents of the HSA-enriched protein fraction. Acronyms: QC - quality control; HSA - human serum albumin; MS - mass-spectrometry; LCa - non-small cell lung cancer; ALB - gene correspondent to the human serum albumin; HP - gene correspondent to the haptoglobin; ORM1 - gene correspondent to the alpha-1-acid glycoprotein.

In summary, in this section, a revisited version of Cohn's method is presented step by step for the depletion of the most abundant proteins in human blood serum and plasma. The fractionation protocol is applied on human serum QC replicas, where it shows to increase the technical noise of the FT-IR spectra compared to full serum samples. However, the noise introduced by the fractionation procedure is lower than the between-person spectral variability as shown via the FT-IR measurements of the human blood full and fractionated plasma of 40 individuals and confirmed via the mass spectrometry measurements of the same samples. The proposed protocol is one of the few methods suited for spectroscopy-based studies because it does not permanently introduce unwanted chemicals in the sample. Mass-spectrometry has helped to address both the performance and the molecular efficiency and make-up of the protein fractions. In particular, the fractionation protocol allows the separation of less abundant proteins and metabolites from three of the most abundant proteins which potentially mask the signatures due to the molecular changes of the less abundant biomolecules in the FT-IR spectra of full biofluids. The protocol is ultimately proved to be efficient, reproducible and suited for both human blood plasma and serum samples. The benefits of combining fractionation and IR fingerprinting are explored in pilot studies for the classification of three cancer entities in the next section.

5.2 Characterization of human blood biofluids via IR spectroscopy and chemical fractionation

Human biofluids cover a wide dynamic range of concentration and the signature of highly abundant molecules can cover the one from the low abundant ones. A chemical fractionation protocol has been proposed to separate the biomolecules into three fractions and the efficiency and reproducibility have been addressed in the previous section. The advantages of combining IR spectroscopy with chemical fractionation are here tested on the human blood serum samples of 334 individuals to address the molecular nature of the spectral variations induced by three cancer entities (Tables 5.1, 5.2 and 5.3): non-small-cell lung cancer, which counts both adenocarcinoma and squamous carcinoma patients, are classified with lung hamartoma patients (a benign condition), chronic obstructive pulmonary disease (COPD) and non-symptomatic individuals (controls); breast cancer cases are classified with non-symptomatic individuals; prostate cancer cases are classified with benign prostatic hyperplasia patients (BPH) and non-symptomatic individuals. The average age and gender of the controls are matched to the corresponding cases, while the smoking status is matched only for the lung cancer cohorts.

Lung cancer				
Cohort	n. individuals	Age	Males/Females	Smokers / not active smokers
Lung cancer	53	68.8 ± 10.2	0.7	0.9
Hamartoma	34	62.7 ± 14.4	0.8	0.9
COPD	26	61.9 ± 15	0.5	0.2
Controls	31	58.6 ± 11.4	0.8	0.7

Table 5.1: Lung cancer detection via chemical fractionation and IR spectroscopy of human blood serum: description of cases and controls. Acronyms: COPD - chronic obstructive pulmonary disease.

Breast cancer		
Cohort	n. individuals	Age
Breast cancer	41	63.7 ± 11.7
Controls	39	63 ± 11.5

Table 5.2: Breast cancer detection via chemical fractionation and IR spectroscopy of human blood serum: description of cases and controls.

Prostate cancer		
Cohort	n. individuals	Age
Prostate cancer	36	60.5 ± 11.2
BPH	38	62.4 ± 10.4
Controls	35	59.7 ± 12.9

Table 5.3: Prostate cancer detection via chemical fractionation and IR spectroscopy of human blood serum: description of cases and controls. Acronyms: BPH - benign prostatic hyperplasia.

Both FT-IR and FRS spectroscopy are applied on full human serum samples as well as on the three corresponding fractions. The advantage of fractionation is that their IR spectra help to address the molecular origin of the fingerprints recorded for full serum samples. The FT-IR and FRS data are first analyzed for the control classes (individuals without cancer) to gain a molecular understanding of the spectral signature of gender. SVM classifications are then reported for the full serum samples and the three fractions to address their role in the spectral signature of lung, breast and prostate cancer.

5.2.1 FT-IR fingerprinting of individuals without cancer

Before looking at the binary classification of cancer, the FT-IR and FRS data of the full serum and the corresponding fractions are first analyzed for the control classes, namely for all individuals without cancer, to investigate the role of each fraction in the binary classification of gender and address the potential of fractionation in combination with FT-IR fingerprinting. Additionally, the IR spectra of the serum samples of the 181 benign and non-symptomatic individuals considered are inspected with the spectra of the plasma samples of the 20 controls analyzed in section 5.1) to qualitatively compare the molecular information carried by the two biofluids.

The SVM binary classification of gender on the FT-IR spectra of the full serum samples and the corresponding three fractions of the 181 controls shows that the metabolite and the albumin-depleted protein fractions are the most important in the classification of males and females. In particular, both fractions return an AUC around 73%, while the albumin-enriched protein fraction returns no contribution to the classification (Figure 5.3), a conclusion that can be robustly achieved only by splitting the different contributions. The full biofluids deliver higher classification efficiencies than the fractions because they carry the additional information of the interaction between the molecules separated in the three fractions and, more importantly, they preserve the information of the residual insoluble pellet lost during fractionation. However, despite the fractions return lower AUCs, they provide a deeper molecular understanding of the signatures observed in full serum.

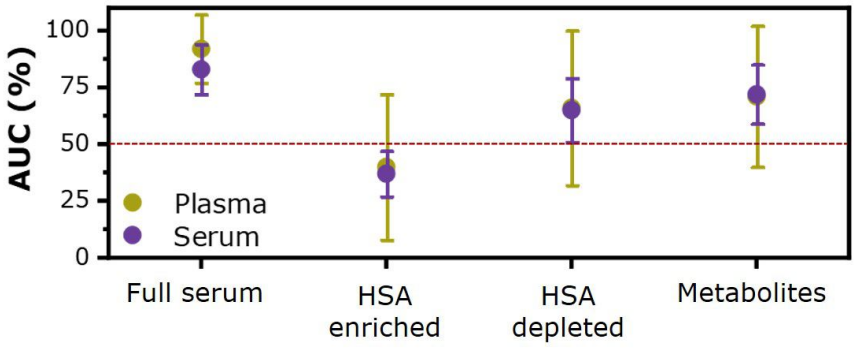

Figure 5.3: SVM binary classification of gender on the FT-IR spectra of full serum and the three fractions of individuals without cancer. The metabolite and albumin-depleted protein fractions return comparable AUCs, while the HSA-enriched fraction does not contribute to the classification. Acronyms: SVM - support vector machine; AUC - area under the curve; HSA - human serum albumin.

The analysis of the fractions of human blood plasma and serum allows to better address the molecular origin of the different IR spectra of these biofluids (Figure 5.4). However, since the individuals of the plasma and serum cohorts are different, this comparison is just qualitative. The FT-IR spectra of full plasma and serum show small differences. The spectra of the albumin-enriched and the albumin-depleted fractions are unexpectedly comparable for the two biofluids. The main difference is due to the presence of 4K-EDTA salt in the tubes used to draw the blood, an anticoagulant agent that forms stable complexes with calcium ions preventing it from interacting with the proteins responsible for coagulation, such as fibrinogen [194]. As a result, plasma retains a larger amount of coagulation proteins like fibrinogen, almost completely missing in serum [195]. The process affects also the composition of metabolites [128]. The high similarity found here between the two biofluids can be due to the low water solubility of the key proteins present in different amounts in serum and plasma, which is the case for fibrinogen, which can potentially be lost in the insoluble pellet during fractionation. From this comparison, it can be deduced that the metabolite fraction retains the largest difference between the two biofluids, in particular, due to the presence in the plasma samples of the EDTA complexes with the metal ions of the biofluids [196].

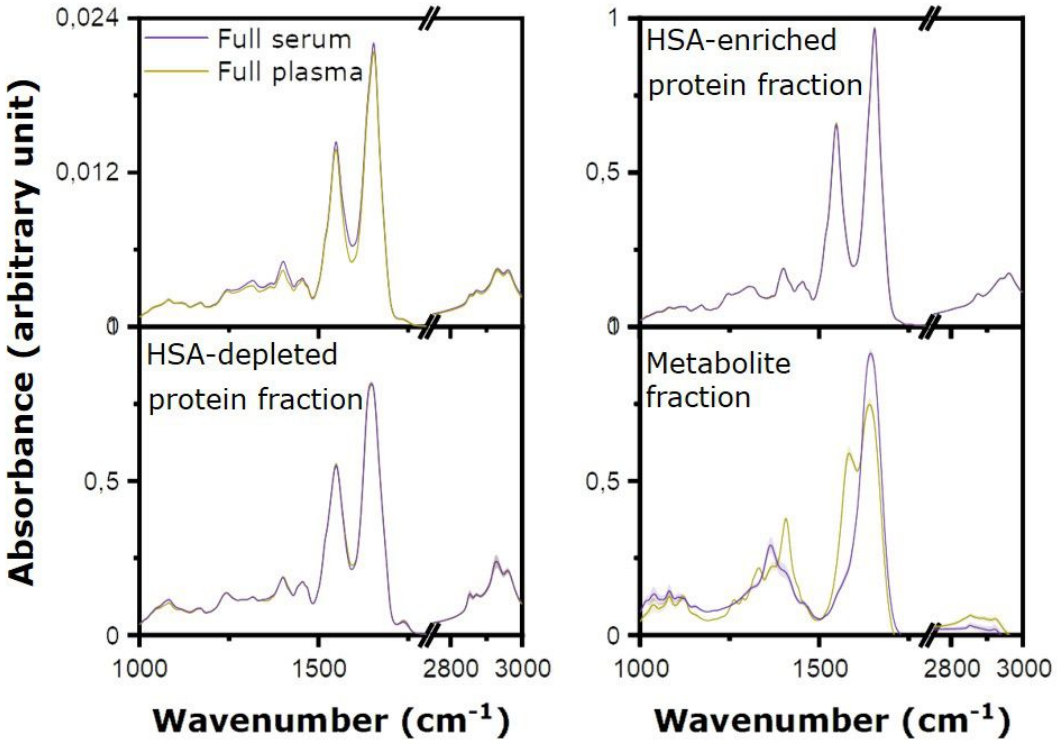

Figure 5.4: FT-IR spectra of full serum and plasma and the respective fractions. The spectra of full serum and plasma are not scaled, while the spectra of the three fractions are all scaled for the same factor in each panel. The spectra of full serum and plasma, as well as of the respective HSA-depleted and HSA-enriched protein fractions, are similar between the two biofluids. The main spectral difference is found in the metabolite fraction and it is due to the presence of EDTA complexes with the metal ions present in the plasma samples. Acronyms: HSA - human serum albumin; EDTA - ethylenediaminetetraacetic acid.

It has been observed that the largest between-person spectral variability associated with the albumin-enriched fraction (Figure 5.1c, d) is in agreement with the high variability of the protein signatures observed for the KORA-FF4 individuals (section 3.2.2 and 4.1) affected by age and inflammation and addressed by the first two principal components. The binary classifications of gender on individuals without cancer via the FT-IR spectra of the three fractions show that the albumin-enriched fraction does not contribute to the classification and that the albumin-depleted protein fraction and the metabolite fraction return similar AUCs. This agrees with the observation that, for the KORA-FF4 population, gender does not correlate with PC1 and PC2, but rather with the principal components addressing the spectral signatures at lower frequencies (PC4 and PC5, section 4.1), which can be attributed to carbohydrates, expected to be in the metabolite fraction, or protein's glycosylation, expected to contribute to the signal of the albumin-depleted protein fraction. This observation suggests that similar conclusions achieved via chemical fractionation can be obtained using PCA to disentangle the contributions of different classes of biomolecules in each component. The analysis of the three fractions has ultimately allowed the comparison of the signatures of human blood serum and plasma, highlighting that EDTA is responsible for the main spectral difference between the two biofluids, which constitute the main difference between the two biofluids in the detection of cancer via IR spectroscopy, addressed in the next section.

5.2.2 FT-IR fingerprinting for cancer detection

Breast cancer and prostate cancer are the most common non-cutaneous cancer in women and men worldwide [197, 198] and non-small-cell lung cancer is among the highest cancer-related cause of deaths in both genders [199]. Several research studies have been trying to identify blood biomarkers for their timely diagnosis [200–209]. In parallel, the research is growing around the use of vibrational spectroscopy of blood biofluids for the detection of non-small-cell lung cancer [117, 210], breast cancer [17, 211, 212] and prostate cancer [213] since such a simple method would be ideal for the time and cost-efficient cancer detection in clinical settings [214]. In this section, a pilot study based on non-small-cell lung cancer, breast cancer and prostate cancer is presented to address the molecular nature of the FT-IR fingerprints of the three cancer entities in human blood serum via applying the fractionation protocol discussed.

The cases and controls analyzed for the three cancer entities are listed in Tables 5.1, 5.2 and 5.3. To start with, the FT-IR signature of non-small-cell lung cancer in human blood serum is first addressed and compared with the spectral signature found in human blood plasma in the pilot study used to characterize the efficiency of the fractionation protocol (section 5.1). Figure 5.1b shows the absorption spectra of non-small-cell lung cancer in full and fractionated human blood plasma and Figure 5.5 shows the corresponding differential fingerprints, defined as the difference between the average spectrum of the 20 cases and the average spectrum of the 20 controls. The differential fingerprints of non-small-cell lung cancer show that the signature in full plasma samples is the result of the lower concentration of albumin and the higher concentration of the proteins in the albumin-depleted fraction for the cases compared to the controls, in agreement with what reported in the literature [215–217].

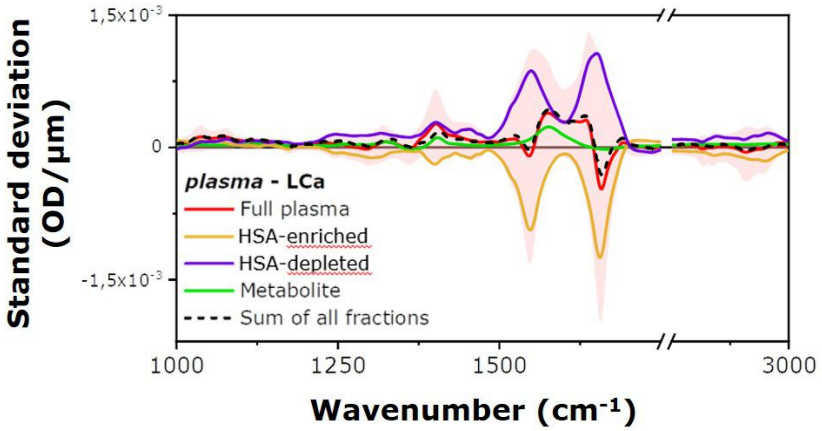

Figure 5.5: Differential fingerprints of LCa in the FT-IR spectra of full human blood plasma samples and corresponding fractions. The differential fingerprints are calculated for the 20 cases and respective 20 controls described in section 5.1 and show that the concentration of HSA is lower while the concentration of the proteins in the HSA-depleted fraction is higher for the cases compared to the controls. The red shaded area shows the standard deviation of all cases in full plasma. Acronyms: LCa - non-small-cell lung cancer; HSA - human serum albumin.

The average AUCs obtained for the classification of non-small-cell lung cancer are comparably high in plasma and serum and have smaller standard deviations in serum compared to plasma because of the higher number of cases (Figure 5.6a). The AUCs of the fractions show that the main signature of non-small-cell lung cancer is due to the albumin-enriched protein fraction, which returns the same AUC as in full serum. Slightly smaller AUCs are recorded for the albumin-depleted protein fraction. The metabolite fractions return significantly different AUCs for the two biofluids: 60% in serum and 90% in plasma. The main difference is due to the presence of EDTA in the plasma samples, as can be observed in Figure 5.4. Even though EDTA is dissolved from the coating of the tubes used during the blood donation of the plasma samples, it is significantly larger in the metabolite fraction of non-small-cell lung cancer patients suggesting a higher concentration of ions that can interact with this molecule to form stable complexes, thus increasing its solubility in the biofluid. In particular, the calcium ions have been reported in the literature to be present in higher concentrations in the blood biofluids of several cancer entities, among which non-small-cell lung cancer, breast cancer and prostate cancer [218]. Therefore, a higher classification efficiency of the metabolite fraction obtained for the human blood plasma compared to serum can be expected also for other cancer entities.

The serum-based cohorts are large enough to compare the classification of non-small-cell lung cancer with non-symptomatic and benign individuals (namely, affected by Hamartomas and COPD conditions) separately. The main difference expected between the two control groups is the potentially high inflammation levels of benign individuals. In particular, for the individuals considered, the highest serum CRP concentration is 2 mg/L for healthy individuals, 62.5 mg/L for individuals affected by Hamartomas and COPD conditions and 183.2 mg/L for non-small-cell lung cancer patients. The individuals affected by Hamartomas and COPD conditions (the benign controls) are more appropriate controls for the classification of cancer because they present symptoms and physiological conditions closer to the ones of the target condition, essential for clinic applications. As a consequence of the higher inflammation levels

due to the benign condition, the AUCs obtained for the classification of non-small-cell lung cancer patients with the benign individuals are lower compared to the AUCs obtained for the classifications with non-symptomatic individuals (Figure 5.6b). In particular, the three main proteins identified in the albumin-enriched fraction, namely albumin, haptoglobin and alpha-1-acid glycoprotein (section 5.1), are associated with acute inflammation [219–221]. Therefore, the albumin-enriched protein fraction encodes for the inflammatory effects of non-small-cell lung cancer and it is, indeed, the only fraction returning a different AUC for the classifications with benign controls compared to non-symptomatic individuals. Therefore, this study shows that the depletion of the albumin-enriched protein fraction can be beneficial to remove the non-specific inflammatory response and isolate a more specific signature of the non-small-cell lung cancer condition. Ultimately, this study shows that the combination of fractionation, FT-IR spectroscopy and mass-spectrometry gives access to the molecular origin of non-small-cell lung cancer spectral signature which can be connected with at least 12 of the most abundant proteins [62], quickly and efficiently detected via infrared fingerprinting.

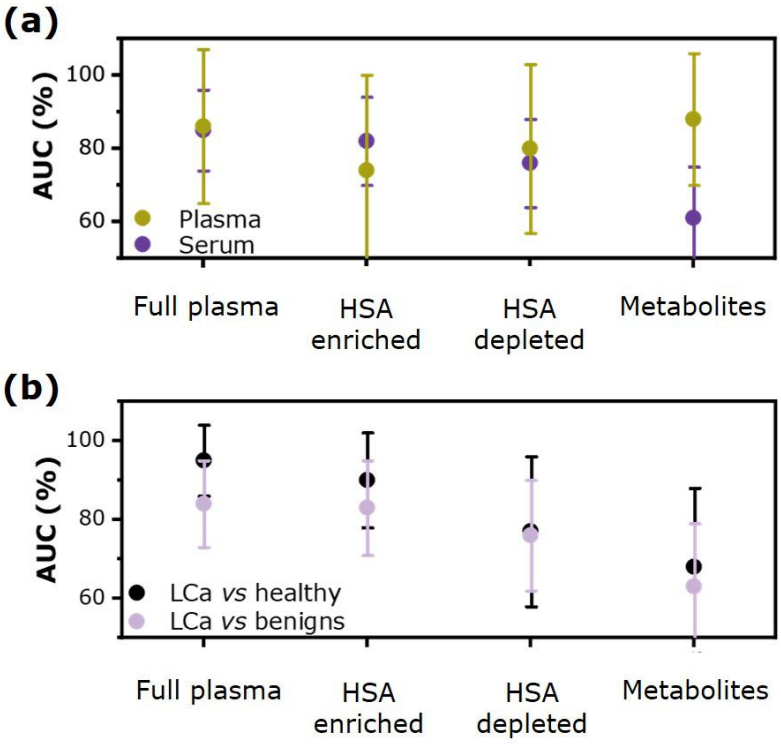

Figure 5.6: SVM binary classification of LCa on the FT-IR spectra of full serum and plasma and the three respective fractions. (a) The AUCs are reported for the classification of LCa with non-symptomatic and benign controls in full human blood plasma and serum and the respective fractions. The main difference between the two biofluids is in the metabolite fractions which return average AUCs of bout 60% and 90% in serum and plasma respectively. (b) The AUCs are reported for the classification of LCa with non-symptomatic and benign controls separately for the FT-IR spectra of full human blood serum and the respective fractions. The main difference is encoded in the HSA-enriched protein fraction which is affected by the inflammatory response of LCa. Acronyms: SVM - support vector machine; AUC - area under the curve; LCa - non-small-cell lung cancer; HSA - human serum albumin.

The serum samples of individuals affected by breast and prostate cancer are analyzed in the same way as non-small-cell lung cancer and the results found for the three cancer entities are compared. The differential fingerprints in full plasma show a higher IR signature for non-small-cell lung cancer compared to breast and prostate cancer (Figure 5.7a).

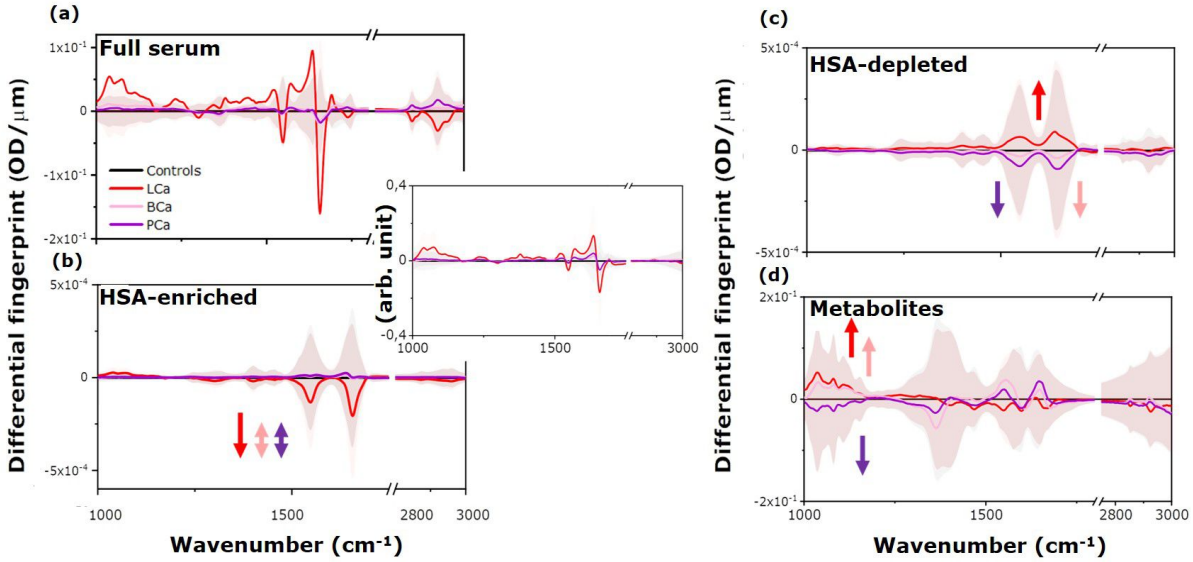

Figure 5.7: Differential fingerprints of non-small-cell lung, breast and prostate cancer for the FT-IR spectra of full human blood serum and corresponding fractions. (a) The differential fingerprints of the normalized absorption spectra of full serum are reported for the three cancer entities calculated with respect to the average spectra of the corresponding non-symptomatic controls. The signatures are different for the three cancer entities showing that the IR features of LCa cases are stronger than for BCa and PCa relatively to their respective controls. (b) The differential fingerprints of not-normalized FT-IR spectra of the HSA-enriched fractions (inset: normalized), (c) the HSA-depleted fractions and (d) the metabolite fractions of the three cancer entities calculated with respect to the average spectra of the corresponding non-symptomatic controls. The relative concentration of proteins and metabolites in the three fractions compared to the respective controls are different for the three cancer entities (arrows). The normalized spectra of the HSA-enriched protein fractions return comparable differential fingerprints for the three cancer entities with the same signatures found for inflammation in section 4.1, highlighting that this fraction carries the general inflammatory response of cancer. The black areas are the standard deviations of all non-symptomatic controls together and the red areas are the standard deviations of all the cancer cases. Acronyms: LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; HSA - human serum albumin.

The differential fingerprints calculated from not-normalized FT-IR spectra show the effect of different cancer types on the total concentration of the biomolecules in each fraction. In particular, the albumin-enriched fractions have small signatures for both breast and prostate cancer compared to non-small-cell lung cancer, which shows the lower concentration of albumin in the cases compared to the controls (Figure 5.7b). Normalizing the spectra before the calculation of the differential fingerprints reduces the impact of absolute concentrations in favor of relative abundances and, for the albumin-enriched protein fraction, returns comparable features for the three cancer entities (Figure 5.7b, inset). The signature of the albumin-enriched protein fraction is the same attributed to the albumin-to-globulin ratio (AGR) typical of inflammation (section

4.1, Figure 3.10b), which confirms the connection of the albumin-enriched protein fraction with inflammation seen in the classification of non-small-cell lung cancer with non-symptomatic and benign controls separately (Figure 5.6).

The albumin-depleted fractions of breast and prostate cancer show an opposite trend compared to non-small-cell lung cancer, showing an overall lower protein concentration in the cases compared to the controls (Figure 5.7c). The differential fingerprints of the metabolite fractions have comparable intensity in the carbohydrate and protein glycosylation region (1000 - 1250 cm^{-1}), showing higher concentration in the plasma samples of cases compared to the controls for non-small-cell lung and breast cancer, with an opposite trend for prostate cancer (Figure 5.7d). Both breast and prostate cancer have comparable signatures at longer wavenumbers in the FT-IR signature of the corresponding metabolite fraction. The IR analysis of the fractions returns a better molecular understanding of the IR signature of the full plasma and serum samples, highlighting the importance of inflammation and the different concentration of albumin, the other serum proteins and of metabolites between the cancer patients and controls in the three cancer entities. The classification efficiencies are higher for non-small-cell lung cancer compared to breast and prostate cancer for the full serum and the three fractions (Figure 5.8), highlighting that IR fingerprinting is potentially better for the detection of this cancer entity. All the average AUCs found for prostate cancer are slightly higher than the values found for breast cancer, but the standard deviations are too high to draw robust conclusions for which the analysis of larger cohorts is needed (see section 4.2.3). From the average AUCs, it is expected that the metabolite fraction has an important role compared to the protein fractions in the binary classification of both prostate and breast cancer which, according to what observed for non-small-cell lung cancer (Figure 5.6a), can potentially be better classified in human blood plasma through the signature of the calcium complexes with EDTA.

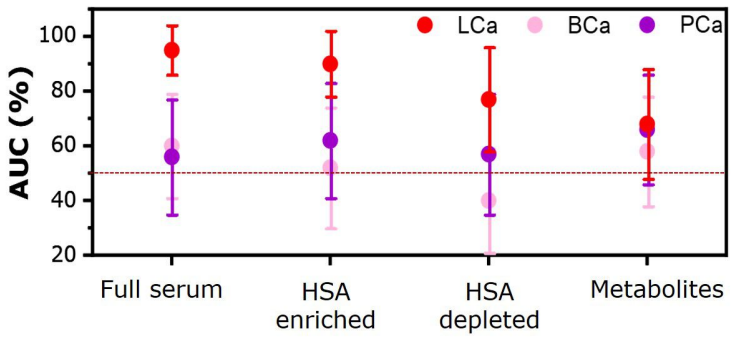

Figure 5.8: SVM binary classification of LCa, BCa and PCa on the FT-IR spectra of full serum and the three respective fractions. The higher AUCs found for LCa for the full serum samples and the corresponding fractions show that FT-IR can potentially perform better in the detection of this cancer entity. The AUCs found for the classifications of PCa are slightly higher than the AUCs found for BCa, but the standard deviations are too large to get robust conclusions and the analysis of more cases is required. From the average AUCs, the role of metabolite fraction is expected to be important for the detection of BCa and PCa, expected to be higher in blood plasma according to what observed for LCa (Figure 5.6a). Acronyms: SVM - support vector machine; AUC - area under the curve; LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; HSA - human serum albumin.

In summary, the pilot studies on the three cancer entities highlight the potential of fractionation for a better molecular understanding of the IR signatures of non-small-cell lung, breast and prostate cancer. The sum of the differential fingerprints of the three fractions reconstruct perfectly the differential fingerprint of full plasma, showing that the additional technical noise introduced by the applied fractionation protocol does not extensively affect the IR signature of the conditions investigates, which agrees with the observation that the technical noise of fractionation is negligible compared with the between-person spectral variability (section 5.1). Therefore, the relevant biological information underneath the signature of non-small-cell lung cancer is retained in the fractions, as it is expected to be the case for all cancer entities also in the serum samples. Chemical fractionation allows comparing the molecular information of non-small-cell lung cancer in human blood serum and plasma showing that plasma can potentially provide higher classification efficiencies because of the signature of the EDTA-ion complexes in the metabolite fractions, expected to be true for all cancer entities for which the concentration of this ion has been reported to be higher compared to the controls [218]. The analysis of the three fractions has allowed to isolate the main proteins affected by the inflammatory response of cancer in the albumin-enriched protein fraction, which returns the highest AUCs for the classification of non-small-cell lung cancer compared to the other fractions as well as to the other cancer entities. For breast cancer, metabolite fraction returns the highest AUC among the fractions, while for prostate cancer the three fractions return comparable AUCs. Because of the important signature of the metabolite fraction in both breast and prostate cancer, their classification efficiencies are expected to be higher in human blood plasma compared to serum, as observed for non-small-cell lung cancer. Despite the fractionation protocol requires about 8 hours thus introducing a time-consuming step in the IR fingerprinting analysis, it provides a better molecular understanding of the IR signature of the full plasma and serum samples showing how the concentration of albumin, the other serum proteins and metabolites change in the target condition compared to the respective controls. Moreover, it allows to chemically separate the proteins mainly connected with inflammation, which make up the albumin-enriched protein fraction, from the other biomolecules thus separating the non-specific signature of inflammation from the more disease-specific molecular changes.

5.2.3 Comparison of FT-IR and FRS fingerprinting for cancer detection

In the previous chapter, the advantages of fractionation for a deeper molecular interpretation of the IR signature of three cancer entities have been shown via FT-IR spectroscopy. Similar to what was done for the common medical conditions and parameters on the KORA-FF4 cohort, the performances of FT-IR are compared with the ones of FRS spectroscopy. In particular, in this section, the data obtained with the current FRS set-up are compared with what reported for the FT-IR spectroscopy for the analysis of the full serum samples and the three corresponding fractions of non-small-cell lung, breast and prostate cancer to address the potential advantages of fractionation in the analysis via FRS spectroscopy.

The analysis presented for the FT-IR spectra in the section above is first performed in the same spectral coverage as FRS (1000 - 1500 cm^{-1}) for the comparison. The AUCs obtained for FT-IR in the reduced spectral coverage are comparable with what was observed for the binary classifications of FT-IR in full-spectra within the standard deviations, showing that most of the

molecular changes can be detected in this spectral range (Figure 5.9). It is, however, probable that the discrepancies between the AUCs found for the reduced and the full spectral coverage are more visible for a larger number of cases because this would reduce the standard deviation of the AUCs.

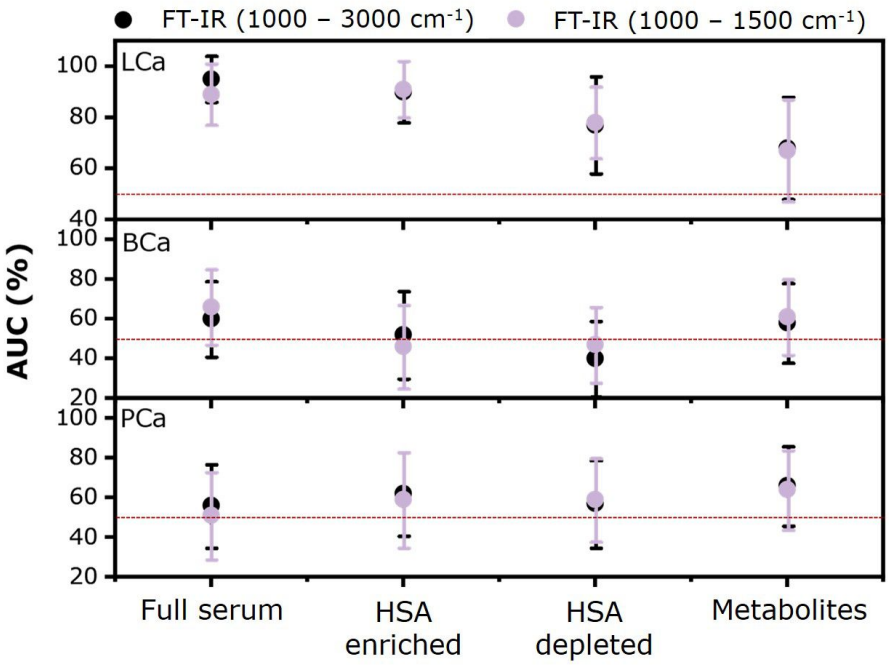

Figure 5.9: SVM binary classification of LCa, BCa and PCa on the FT-IR full and reduced spectral coverage of full serum and the three respective fraction. The AUC obtained from the SVM binary classifications of LCa, BCa and PCa with matched non-symptomatic individuals on the FT-IR spectra of full serum and the three corresponding fractions are comparable within the standard deviation for the reduced (1000 - 1500 cm^{-1}) and the full spectral coverage. Acronyms: SVM - support vector machine; AUC - area under the curve; LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; HSA - human serum albumin.

Before getting into the details of the FRS analysis, it is important to optimize the preprocessing and identify the protocol able to minimize the technical noise, similarly as done in section 3.1.2. The FRS measurements of the three cancer entities have been performed one year before the KORA-FF4 measurements with similar experimental settings but with less diagnostic to monitor the laser fluctuations in real-time and with a worse synchronization between the delay stage and the chopper trigger which results in the walk-off of the time traces measured consecutively. The walk-off is corrected via the *global-T-position* (GT), which centers the maximum values of all reference water traces and the sample time traces to a common zero. The technical noise is measured as the standard deviation of QC replicas (section 3.1.3). For the EMFs of the full serum samples, the Hilbert centering (HC) and the GT transformation are applied to center the data, then the echo correction (EC) is applied to reduce the noise due to the back refection at the EOS crystal around 1.5 *ps* followed by standardization (ST), necessary to compensate for the changes of the exciting laser pulse which affect the molecular signal in the whole time trace (Figure 5.10a). Because of the higher instability of this pilot study, a

different optimal preprocessing is found for each fraction (Figure 5.10b). The technological improvements on the FRS set-up performed during the year between the measurement of this pilot study and the KORA-FF4 samples have been able to provide a more robust and reproducible technique, without which long measurements like the KORA-FF4 (over three months) would have not been possible.

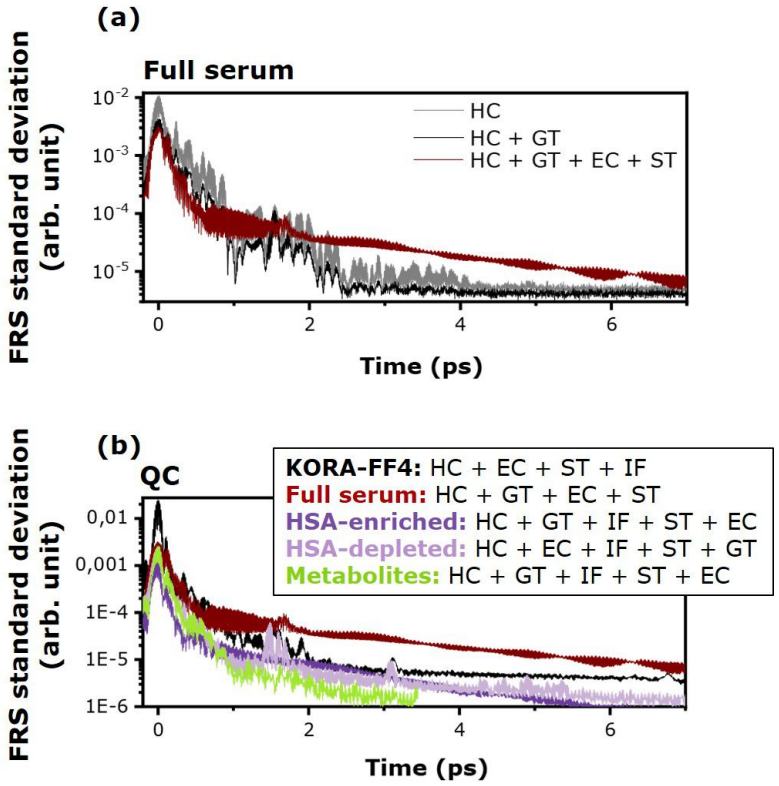

Figure 5.10: Optimal preprocessing for the EMFs of LCa, BCa and PCa serum samples and the three respective fractions. (a) The technical noise, expressed as the standard deviation of the EMFs of QC replicas, is reported for the measurements of full serum for different preprocessing protocols. The HC and GT transformation center the time traces to a common zero; EC reduces the noise due to the back-reflection at the EOS crystal around 1.5 *ps*; ST compensate for the fluctuations of the exciting pulse in the whole temporal range. (b) The technical noise of the EMFs of full serum is compared with the technical noise of the three fractions of QC replicas and the KORA-FF4 cohort after applying the optimal preprocessing, measured one year later in more robust experimental settings. The best preprocessing found and applied on the FRS measurements of each sample type is reported in the legend. The technical noise found for the three fractions is lower compared to the one obtained for the measurements of full serum ad plasma because of the lower chemical complexity of the fractions. The need for different preprocessing protocols and the higher technical noise recorded for the full serum samples compared to the one of the KORA-FF4 FRS measurements show that the experimental settings have improved considerably during the one year between the two measurement campaigns, necessary for long measurements like the KORA-FF4 campaign. Acronyms: EMF - electric-field-resolved molecular fingerprint; LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; QC - quality control; HC - Hilbert centering; GT - global-T-position; EC - echo-correction; ST - standardization; IF - interference correction; EOS - electro-optic sampling; HSA - human serum albumin.

The technical noise found for the FRS measurement of the fractions is lower compared to the technical noise found for the KORA-FF4 measurements after the optimal preprocessing (section 3.1.2) because of the lower chemical complexity (Figure 5.10b). However, the technical noise found for full serum after applying the optimized preprocessing is higher from 1 *ps* to the end of the time trace compared to the technical noise found for the KORA-FF4 measurements. The noise at longer time found for the FRS measurements of full serum increase after applying standardization, which is based on the ratio between the EMFs of the water reference and are, therefore, more prone to errors at long times where the signal gets smaller. A similar issue is observed for the KORA-FF4 FRS measurements (Figure 3.6c), however, while for the KORA-FF4 measurements applying the EC before ST helped to reduce the technical noise at longer times, this does not happen for the measurement of these samples. The reason for this is currently unclear.

Despite the higher technical errors introduced by standardization, this is a necessary step for the robust analysis of the EMFs, as shown by the binary classifications of the EMFs on full serum sample for the detection of prostate cancer with two preprocessing protocols: one based only on the centering transformations (HC and GT) and one with both EC and ST (Figure 5.11). The comparison of the SVM performances along the time trace for the two preprocessing protocols shows that standardization reduces the overfitting rate in the whole time window, calculated as the distance between the average AUC of the training set and the standard deviation of the test set, compared to applying only the centering options.

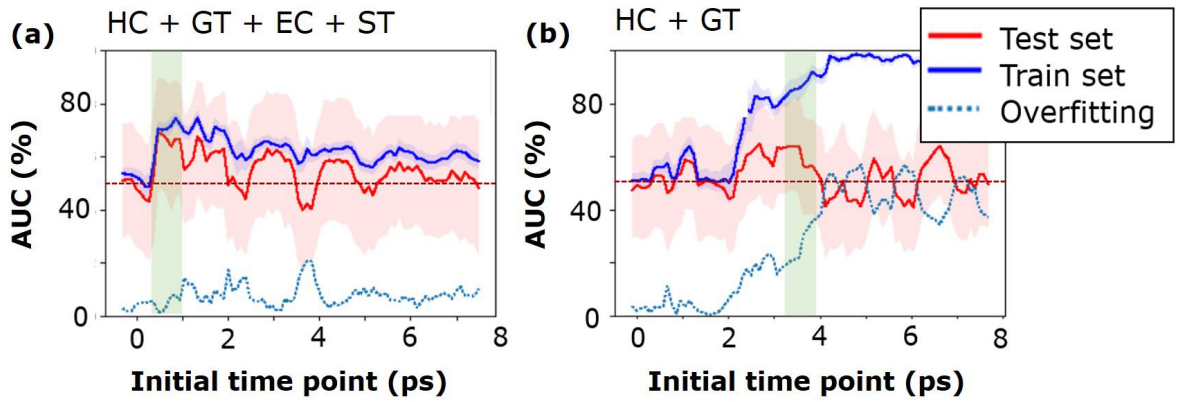

Figure 5.11: SVM binary classification of PCa on the EMFs of full serum for different preprocessing protocols. (a) The AUCs in time are more robust applying HC, GT, EC and ST and have lower overfitting rates than (b) the AUCs obtained applying only the centering transformations (HC and GT). The red dashed lines show the AUCs of 50% and the green shaded areas highlight the best time window. Acronyms: SVM - support vector machine; AUC - area under the curve; EMF - electric-field-resolved molecular fingerprint; PCa - prostate cancer; HC - Hilbert centering; GT - global-T-position; EC - echo-correction; ST - standardization; IF - interference correction.

As explained in section 2.3.2 and done for the analysis of the FRS data in the previous chapter, the SVM binary classifications are performed using sliding windows to address the one providing the highest classification rate, identifies as the optimal time window. This procedure provides a temporal resolution of the AUCs. In particular, the AUCs found for non-small-cell lung cancer are comparable between the FT-IR and FRS data (Figure 5.12). The average AUCs

found for breast cancer are higher for the EMFs traces of both protein fractions, going from 50% for FT-IR up to 65 - 70% for FRS. Because of the higher AUCs found for FRS in the protein fractions, the similar AUCs found for FT-IR and FRS in full serum are unexpected and can potentially be the consequence of the higher technical noise recorded at long times for the EMFs of full serum. The FRS classification efficiencies are higher compared to FT-IR also for the classification of prostate cancer in full serum, for which FT-IR returns no separation and FRS returns an AUC of about 70%, in agreement with the higher average AUCs found for the two protein fractions. Therefore, despite the instability of the FRS set-up employed for these measurements, the method is shown to perform already better than FT-IR spectroscopy for breast and prostate cancer.

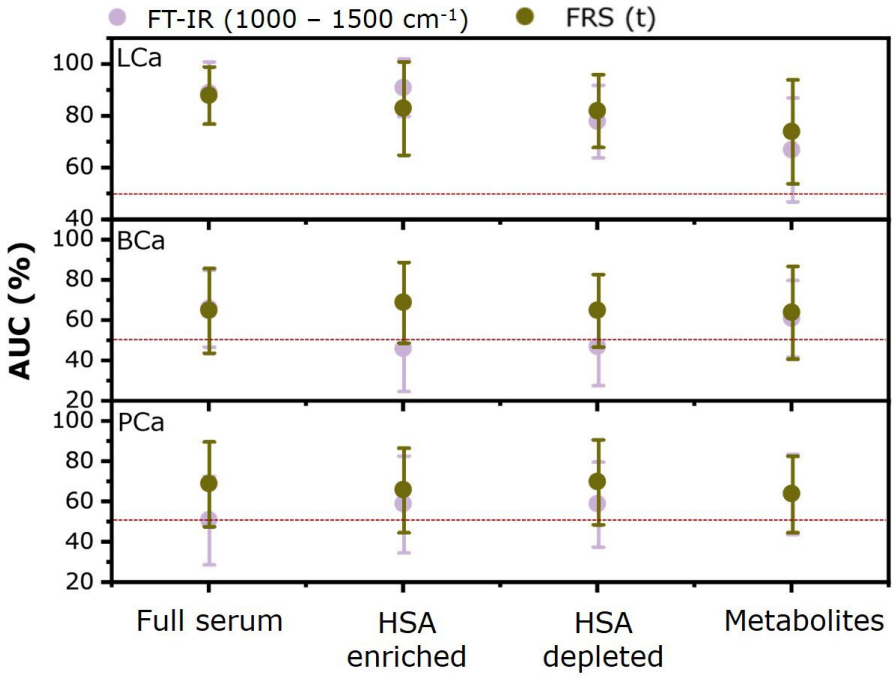

Figure 5.12: SVM binary classification of LCa, BCa and PCa on the FT-IR and FRS data of full serum and the three respective fractions. The AUCs found for FT-IR are obtained for the same spectral range covered by FRS (1000 - 1500 cm^{-1}) and are compared with the AUCs obtained for the FRS in the time domain for the classifications with matched non-symptomatic individuals. The AUCs found for FT-IR and FRS are comparable for the classification of LCa. FRS returns higher AUCs for the classification of BCa and PCa in the two protein fractions, which explain the higher AUC found for FRS also in the full serum for the classification of PCa, while the AUC in full serum for breast cancer is unexpectedly the same as found for the FT-IR data. Acronyms: SVM - support vector machine; AUC - area under the curve; FRS - field-resolved spectroscopy; LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; QC - quality control; HSA - human serum albumin.

Despite the trends of the AUCs in time are noisy because of the small number of cases, the overfitting rates are generally low below 2 ps and in the optimal time windows identified for each condition (Figure 5.13). For non-small-cell lung cancer, the optimal time windows found in full serum and the three respective fractions are around 1 ps. Therefore, the biological information of this cancer type is confined in time.

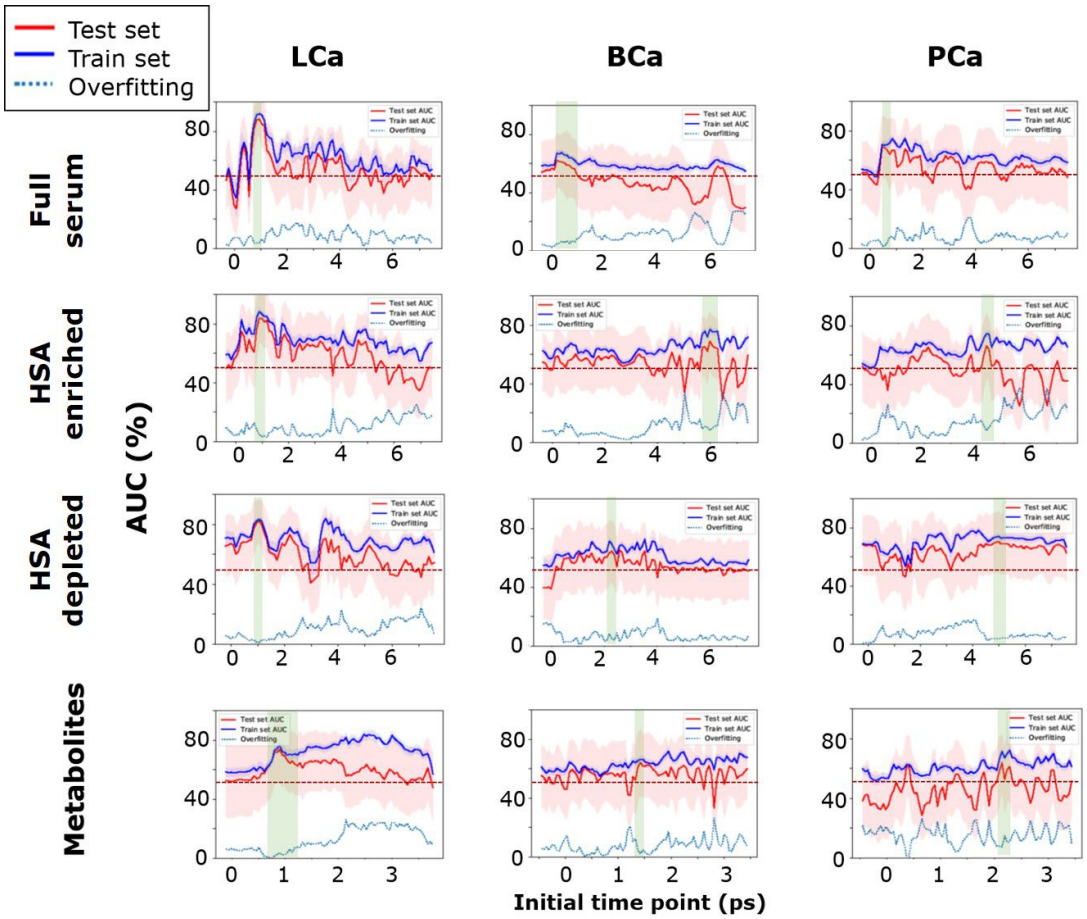

Figure 5.13: SVM binary classification of LCa, BCa and PCa on the EMFs of full serum and the three respective fractions. The trends of the AUCs in time are reported for the classifications with matched non-symptomatic controls. The green shaded areas highlight the optimal temporal windows. The classifications of LCa return similar optimal time windows, namely the temporal windows for which the AUC reaches the maximum value for that classification, are around 1 *ps* for the full serum and the respective fractions. The optimal temporal windows found for BCa and PCa are different for full serum and the three respective fractions. However, the AUC trend in time found for PCa reaches comparably high values in the whole time trace, thus covering also the temporal windows of the fractions. For BCa, the AUC found for full serum is affected by higher overfitting rates, probably due to the high technical noise found for these samples (Figure 5.10b), which can explain why the AUCs found for FRS are higher compared to FT-IR for the two protein fractions and not for full serum. Acronyms: SVM - support vector machine; AUC - area under the curve; EMF - electric-field-resolved molecular fingerprint; FRS - field-resolved spectroscopy; LCa - non-small-cell lung cancer; BCa - breast cancer; PCa - prostate cancer; HSA - human serum albumin.

For breast cancer, the optimal time windows found for full serum and the individual fractions are different. In particular, the albumin-depleted protein fraction and the metabolite fraction are stable to overfitting in the whole time trace and return high AUCs from 0 to 4 *ps* and in the whole time window respectively. The optimal time window found for the classification of breast cancer in full serum is centered around 1 *ps*, mostly because of the higher overfitting rates reached at longer times. However, at 6.2 *ps*, corresponding to the optimal temporal window

found for the albumin-enriched protein fraction, the overfitting of full serum decreases reaching AUCs comparable to the maximum. The classification of breast cancer in full serum shows high overfitting rates that can be the consequence of the higher technical noise at long times due to standardization (Figure 5.10b) and explain the similar AUCs found for FT-IR and FRS despite the higher classification efficiencies found for the two protein fractions. The optimal time windows found for the classification of prostate cancer are also different for full serum and or the three fractions. However, the AUCs trend in time for full serum reached comparably high values at all times with low overfitting rates, highlighting that the three fractions contribute in different temporal windows to the classification of prostate cancer with comparable AUCs.

In summary, the classification efficiencies found for the FRS and FT-IR data are comparable for non-small-cell lung cancer. However, FRS performs better than FT-IR for the detection of breast and prostate cancer because of the higher AUCs identified in the albumin-enriched and albumin depleted protein fraction of the two cancer entities, reflected also in the higher classification efficiency of the prostate cancer in full serum via FRS compared to FT-IR. Despite the low number of cases, the optimal temporal windows have low overfitting rates, thus ensuring the robustness of the classifications. Ultimately, the combination of FRS with the fractionation protocol allows identifying the contribution of different groups of biomolecules to the classification efficiency, in particular showing that the signature of the two protein fractions are more important in the binary classification of breast and prostate cancer for the FRS data than for FT-IR, and to address their contributions at different times. The classification rates reported in this study are lower compared to other results reported in the literature for similar analysis [17, 117, 210–213]. However, the study discussed here is based on a very small number of cases and controls and it is expected to provide more robust results as well as higher classification efficiencies in larger studies. The higher classification rates obtained for FRS compared to FT-IR in the same experimental settings is impressive considering the limitations of the technique at the moment of the measurements and show that FRS is potentially a better tool for the detection of breast and prostate cancer.

5.3 Concluding remarks

A re-adaptation of Cohn's method is presented and adopted for the fractionation of human blood serum and plasma samples in a metabolite fraction and two protein fractions: the albumin-enriched protein fraction, which is constituted by about 85% of the human serum albumin and other two of the most abundant proteins, and the albumin-depleted protein fraction constituted by the remaining serum proteins. The technical noise due to the chemical fractionation procedure is lower than the between-person spectral variability of the FT-IR spectra of human blood plasma, also confirmed via mass-spectrometry analysis which has been used to address the molecular composition of the protein fractions. Chemical fractionation via precipitation is among the few techniques for protein depletion compatible with complex samples like human blood serum and plasma as well as with spectroscopic applications since it does not introduce unwanted chemicals in the samples and does not affect the chemical bonds of the analytes. However, the adopted protocol requires about 9 hours for the fractionation of 24 samples, but it can be scaled up by using equipment able to handle more samples and it can be speed-up by employing robotic liquid-handling systems. The main drawback of the chemical precipitation-based protocols is that a small part of the sample cannot be dissolved

back in water, as is the case for a pellet of the albumin-depleted protein fraction. However, this has been estimated to be a small percentage of the albumin-depleted protein fraction, which is confirmed by the observation that the sum of the average spectra of the three fractions reconstruct the average spectrum of the full blood biofluids.

Chemical fractionation combined with IR fingerprinting has been used to address the molecular nature of the spectral signature of gender among healthy individuals highlighting that the metabolite fraction and the albumin-depleted protein fraction return the highest classification efficiencies, in agreement with the correlation found for the analysis of the KORA-FF4 cohort between gender and the spectral signatures at lower frequencies due to the vibrations of carbohydrates, expected to be in the metabolite fraction, and protein's glycosylation, expected to be in the albumin-depleted protein fraction (section 4.1).

The three main proteins that constitute the albumin-enriched fraction, in particular albumin, haptoglobin and alpha-1-acid glycoprotein, are connected with inflammation [219–221]. The spectral signature of these proteins is indeed similar to the IR fingerprint found for high CRP values, a marker of inflammation, in the KORA-FF4 cohort (section 3.2.2). Moreover, the high between-person variability addressed by this fraction is in agreement with the high impact of inflammation on the between-person spectral variability observed in the previous chapters (section 3.2.2 and 4.1). The albumin-enriched protein fraction returns the highest AUCs for the classification of non-small-cell lung cancer with respect to the other fractions highlighting the importance of this unspecific signature in the analysis of this cancer entity. The chemical fractionation protocol can be adapted to deplete only the albumin-enriched protein fraction to reduce the concentration of the most abundant serum protein, albumin, together with the main proteins connected with the non-specific inflammatory response of the body.

The FT-IR analysis of the full serum samples and the corresponding fractions shows that the main signature of breast cancer comes from metabolites, while for prostate cancer the three fractions return similar classification efficiencies thus highlighting a different molecular nature of the signature of each cancer entity underneath the IR spectra of the full serum samples. Moreover, the comparison of the classification efficiency for non-small-cell lung cancer in human blood serum and plasma shows that the metabolite fraction of plasma returns higher AUCs because of the signature of the EDTA-ion complexes not present in serum, which are expected to be more abundant in the cases because of the higher calcium ion concentration in the blood serum of several cancer entities [218], among which non-small-lung cancer, but also breast and prostate cancer, for which a similar result is expected.

The classification efficiencies found for non-small-lung cancer via FRS are similar to the ones obtained via FT-IR for full serum and the respective fractions. FRS returns higher AUCs compared to FT-IR for the classification of breast and prostate cancer, especially for the albumin-enriched and albumin-depleted protein fractions. Therefore, the combination of FRS analysis with chemical fractionation allows identifying the molecular origin of the better performances of FRS compared to FT-IR in the two protein fractions. The robustness of the results, despite the low number of cases, is supported by the low overfitting rates found in the optimal temporal windows of each classification. However, the best classification efficiencies that can be achieved via IR spectroscopy for these cancer entities are expected to be higher and can be achieved by analyzing a larger number of cases (section 4.2.3). Higher classification rates have been already reported in the literature for similar techniques [17, 117, 210–213]. A larger study of the considered cancer entities via FRS spectroscopy is planned in the near future. In conclusion,

considering the technical limitations of the technique at the moment of the measurements and the higher technical noise compared to what has been observed for the KORA-FF4 cohort (Figure 3.6c), the better performances of FRS compared to FT-IR are first results of great auspicious for future clinical applications.

In this section, it has been shown that chemical fractionation gives access to a better understanding of the molecular nature of the IR fingerprints of full biofluids and the role of different classes of biomolecules in different conditions. However, the separation in different fractions to this level can potentially be addressed via principal component analysis, able to separate the spectral features connected with different families of biomolecules. Moreover, it does not provide advantages to obtain information about the molecular changes of low abundant molecules, which can potentially provide higher classification efficiencies for many diseases, because the level of chemical complexity, namely the dynamic range of concentration covered, is still too high for the three fractions isolated with the protocol adopted and the signal from low abundant molecules is still not accessible. To isolate the IR signature of low abundant molecules, other fractionation techniques need to be used in combination with infrared spectroscopy, such as chromatographic techniques, which are currently being implemented in our laboratories.

Chapter 6
Conclusions

In this dissertation, the potential of FT-IR and FRS fingerprinting of human blood biofluids for liquid biopsy is addressed via FT-IR and FRS spectroscopy in a cross-sectional population-based cohort, the KORA-FF4, and validated in an independent clinic-based cohort, the L4L. In the first place, the origins of the large between-person spectral variability are investigated among the healthiest individuals of the cross-sectional population-based cohort, which represents a general German population. The spectral signatures and diagnosis efficiencies for several common endpoint and intermediates medical conditions are then identified using both FT-IR and FRS spectroscopy. The potential of combining these spectroscopic methods with a reproducible chemical fractionation protocol is ultimately explored in the frame of cancer diagnosis.

The analysis of the large cross-sectional population-based KORA-FF4 cohort has allowed identifying age and inflammation as the main source of the biological difference between individuals causing the large between-person spectral variability recorded for the IR signatures of lipids (1750 - 3000 cm^{-1}) and proteins (1250 - 1750 cm^{-1}). The influence of age on the lipid content of human blood biofluids was already well-known [93, 94, 99, 100]. Aging and inflammation are connected [97, 98], explaining the impact of age also on the spectral signatures attributed to inflammation [61, 92]. In the spectral coverage of FRS spectroscopy (1000 - 1500 cm^{-1}), gender and inflammation have been identified as the main sources of the between-person spectra variability, affecting the IR signatures attributed to carbohydrates and protein glycosylation. The influence of gender on the spectral signature of carbohydrates can be due to the different glucose tolerance observed in males and females in [106]. Gender and inflammation are both connected with protein glycosylation, as reported in [101–103]. The contribution of these spectral signatures (1000 - 1500 cm^{-1}) to the total variability recorded in full-spectra (1000 - 3000 cm^{-1}) is about 2%, in agreement with the observation reported in the literature that age induces stronger changes in the human blood plasma composition compared to other factors, especially for proteins and lipids [107].

Despite the large between-person spectral variability, IR spectroscopy combined with SVM binary classifier has been proven useful for the efficient detection of several common medical conditions. The comparison between the KORA-FF4 and the L4L cohorts has allowed identifying that matching the symptomatic individuals to the respective controls allows isolating disease-specific IR signatures in very different cohorts by reducing the spectral contributions of non-specific factors that correlate with the medical disease target of the binary classification. Despite matching the controls to the cases reduces the number of controls, thus lowering the statistical power of the analysis [144], it provides access to the specific signature of each disease independently on the study showing the robustness of the approach.

The best performances found for IR fingerprinting have been obtained for diabetes, hypertension and high blood lipids, with promising results also for the detection of stroke and heart attack which can be useful for the timely detection of people at high risk. Infrared spectroscopy has provided low detection efficiencies for COPD, former cancer and asthma conditions. However, the actual efficiency of each classification can be robustly addressed only if the statistical power of the study is high enough [144]. The statistical power increases with the number of individuals analyzed. The minimum number of cases for robust binary classifications has been estimated to be around 130 individuals, in agreement with what has been reported in the literature [150] and are expected to have general validity in similar studies.

The molecular interpretation is beyond the purpose of this study but it is attempted for the medical conditions for which the detection via IR spectroscopy is promising. The spectral signatures at 1750 – 3000 cm^{-1} are mostly due to lipids and are important in the binary classification of high blood lipids, heart attack and stroke. However, the molecular origin of these signatures can be different for each condition and, according to what has been reported in the literature, it could potentially be due to lipoproteins for both high blood lipids [134], with the contributions of triglycerides and glycerol at lower frequencies [140], and stroke [141], with the contribution of amino acids [142] and metal complexes [143] at lower frequencies. The spectral signature of lipids recorded for heart attack can potentially be connected with the low-density lipoprotein (LDL) [132–134], together with contributions at lower frequencies due to the lower fasting blood glucose of individuals who had a heart attack [135]. The signature of hypertension is more difficult to interpret and it could potentially be connected with urea and creatinine, reported as biomarkers of this medical condition [136]. To the best of our knowledge, this is the first study of hypertension via the FT-IR spectra of human blood biofluids. The main spectral features found for diabetes can be attributed to carbohydrates and proteins' glycosylations, in good agreement with the well-known effect of glycemic dysregulation [129, 130], while the spectral signature of prediabetes can have multiple origins. In particular, the IR fingerprint recorded for IGT can potentially be connected with glycine, lysophosphatidylcholine (18:2) and acetylcarnitine which have been reported in the literature as biomarkers and provide a similar AUC as FRS spectroscopy [159]. The IR fingerprint of IFG can be due to glucose and proteins' glycosylations as well as to amino acids and phospholipids, according to [165, 168]. Finally, the signature of IGT/IFG seems a combination of the IR signatures found for the IFG and IGT conditions separately.

The use of MIR spectroscopy for detecting and monitoring the blood content of glucose has been previously suggested by other published studies [169, 170], but our observations, in agreement with other studies [171], suggest that the detection of prediabetes should not rely only on the glycemic dysregulations alone as this can be a late consequence of prediabetes. The ATR-FTIR spectroscopy has been applied for the detection of diabetes [131], showing classification efficiencies similar to what has been obtained in this study, and for the detection of prediabetes [118], where higher classification efficiencies are observed compared to this study. However, the study reported in [118] has been performed on a low number of cases, therefore leading to low statistical power, the three prediabetes entities are classified together, a debatable approach since the three prediabetes conditions have different origins and effect on the composition of blood, and the classifications are performed with unmatched controls and are potentially biased by unspecific factors.

The study presented here is, to the best of our knowledge, the first showing how clinical

parameters connected with diabetes and prediabetes affect the classification efficiencies of IR fingerprints. In particular, it has been shown here that insulin is a worse classifier than IR spectroscopy, in agreement with the lower AUC reported in the literature for the classification of diabetes using only the concentration of insulin [160]. Fasting plasma glucose, as well as HbA1c concentrations that correlate with the blood glucose content, provide classification efficiencies comparable with the ones obtained for FRS for the detection of IGT, for which FT-IR spectroscopy is shown to be a bad diagnostic tool. Higher classification efficiencies compared to the ones obtained in this study have been reported for the classification of IGT via fasting capillary glycemia (FCG) [161], fasting plasma glucose (FPG) and HbA1c [162] as well as via mass-spectrometry [164], but the analysis of a larger number of cases via FRS spectroscopy is expected to provide better classification performances. The classification efficiencies obtained via FRS spectroscopy are similar to the ones reported for the detection of diabetes via FCG [161], FPG and HbA1c [162] and for the detection of IFG via mass-spectrometry [165] highlighting that already at this initial stage, this technique provides good diagnostic efficiency.

The last chapter of the dissertation presents a chemical fractionation protocol to separate the most abundant proteins, among which human serum albumin (about 50-60% of the total serum protein concentration), from the less abundant proteins, as well as to isolate a fraction constituted by metabolites. The chemical fractionation is used to deplete the highly abundant molecules, which have intense spectral signatures, and identify the spectral features of less abundant molecules. Mass-spectrometry has been used to confirm the reproducibility and efficiency of the protocol and to address the molecular composition of the protein fractions. The protocol adopted is compatible with spectroscopic techniques, it introduces a technical noise smaller than the between-person spectral variability among individuals and it can be easily scaled up and speed up by employing robotic liquid-handling systems.

Chemical fractionation combined with IR fingerprinting has been shown to provide a deeper knowledge on the molecular nature of the IR fingerprints of human blood serum and plasma in a pilot study for the detection of non-small-cell lung cancer, breast cancer and prostate cancer. In particular, albumin, haptoglobin and alpha-1-acid glycoprotein are the main proteins of the albumin-enriched protein fraction and are here observed to be connected with inflammation, in agreement with the literature [219–221]. Therefore, fractionation could be implemented to reduce the non-specific IR signatures of inflammation in the analysis of medical conditions, as reported here for the three cancer entities investigated. The classification efficiencies found in this study for non-small-cell lung cancer, breast cancer and prostate cancer are lower compared to what can be found in the literature for similar techniques [17, 117, 210–213], but the limitations can arise by the small number of cases considered in this pilot study. Despite it provides a better molecular understanding of the IR signatures, the fractions obtained with the adopted protocol still cover a large dynamic range of concentration and, therefore, do not fully isolate the signature of low abundant molecules. To reach this goal, IR spectroscopy should be combined with a technique able to provide a deeper fractionation, such as chromatographic techniques. The combination of infrared spectroscopy and high-performance liquid chromatography (HPLC) is currently being implemented in our laboratories.

In summary, it can be concluded that age and inflammation are responsible for the large between-person spectral variability of non-symptomatic and symptomatic individuals, which does not prevent IR spectroscopy combined with SVM classifier to detect common medical conditions in human blood serum and plasma. In particular, IR spectroscopy is promising

for the detection of diabetes, prediabetes, hypertension and high blood lipids, but also for stroke and heart disease, for which larger studies are needed, while it is not expected to be a good classifier for the detection of asthma, COPD and former cancer patients. Different studies reported in the literature want to identify the best biofluids between human blood serum and plasma [114, 127]. In the study reported here, there are no major differences for the classification of common medical conditions, which can be explained based on the similar chemical composition of the two blood biofluids [128], while human blood plasma is expected to perform better for the classification of cancer because of the spectral signature of the complexes of calcium ions, present in higher concentration in several cancer entities [218], with the ETDA used as anti-coagulant during the plasma blood donation. However, the investigation of different cancer entities in larger cohorts in both serum and plasma is required to confirm this conclusion.

The data recorded with the newly developed FRS technique have a dependence on the measurement day which can be reduced but not eliminated by preprocessing. Nonetheless, FRS returns similar or higher classification efficiencies compared to a state-of-the-art FT-IR spectrometer thanks to the lower LOD achievable via FRS [52]. In particular, the classification efficiencies found for the detection of prediabetes, breast and prostate cancer are higher for FRS compared to FT-IR, thus showing that FRS is potentially a better tool for the detection of this condition. Moreover, the resolution in time makes it easier to disentangle the contributions of different factors which would overlap in the frequency domain thus returning very different signatures for each parameter or medical condition investigated and it has been proven useful to isolate disease-specific spectral signatures from the signature of unspecific parameters that correlate with the target disease. Further technical developments are currently being implemented on the FRS set-up to make it more robust to the daily changes as well as to boost its classification efficiency by expanding the spectral coverage and the detection dynamic range.

Besides the technical improvements of the technique, a better understanding of the IR spectral signature of the general population is still needed. To this end, the investigation of other factors not considered in this dissertation, such as the hormonal status, diet and medications, is important as it can potentially provide more information about other sources of between-person variability. The effect of common parameters on the IR spectra of human blood biofluids is indeed useful for the identification of clusters of individuals with similar IR fingerprints with lower between-person variability, potentially beneficial for the detection of medical conditions. Age and inflammation have an important impact on the IR fingerprints, but aging proceeds at different rates and with different mechanisms in each individual [97]. Therefore, the effect of age seen in the IR spectra could reflect the chemical composition of human blood biofluids for specific “effective age” [97] or “ageotypes” [97], a connection that would be extremely interesting to explore.

The study reported in this dissertation on diabetes and prediabetes, as well as all the ones reported in the literature, is based on the results of the OGTT test, which has been reported to have low reproducibility and efficiency [166, 167], thus making it impossible to assess if the miss-classifications of IR spectroscopy are real. Similar issuers are true for all medical conditions and can be potentially overcome by the longitudinal analysis of the KORA individuals via both FT-IR and FRS spectroscopy planned in the near future.

In conclusion, the implementation of IR spectroscopy in clinical settings is very attractive as it offers the possibility to timely and simultaneously detect and monitor different medical

conditions in a fast and non-invasive fashion. This dissertation reports one of the largest studies investigating the efficiency of FT-IR spectroscopy for disease detection in human blood biofluids as well as the first large study using field-resolved spectroscopy showing that IR fingerprinting generally provides good classification efficiencies for several medical conditions. Already at this initial stage, FRS spectroscopy provides better performances for the detection of many diseases compared to FT-IR spectroscopy highlighting that the further developments currently implemented can bring this technique to be the clinic diagnostic tool of the future.

Appendix - The effect size

The effect size of each common parameter and medical condition discussed for the KORA-FF4 cohorts is reported. The effect size is proportional to the classification efficiency (AUC) at each frequency and it is calculated as the ratio between the differential fingerprint, namely the difference between the average spectrum of the controls from the average spectrum of the cases, and the standard deviation of the controls [222]. This approach is based on the assumption that the distribution of cases and controls is Gaussian at each frequency and that the standard deviation of the cases is comparable to the standard deviation of the controls.

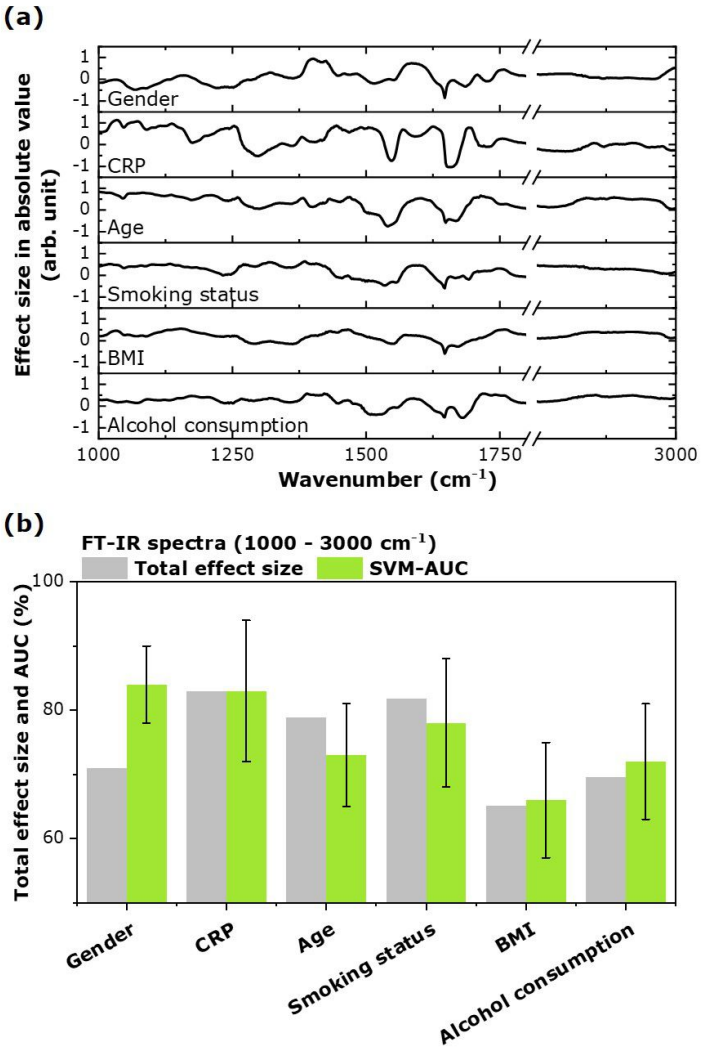

Figure 6.1: Effect size of common parameters among healthy individuals on the FT-IR data of the KORA-FF4 cohort. (a) The effect size of each common parameter shows that different frequencies are relevant for the binary classification of different common factors among the same healthy individuals. (b) The total effect size of each parameter is calculated as the area under the absolute values of the effect size (panel (a)) normalized to bring the highest value (for CRP) to the correspondent AUC. The comparison with the AUCs shows that the two approaches return comparable results, except for gender for which the AUC is higher compared to the total effect size. Acronyms: AUC - area under the curve; CRP - C-reactive protein; BMI - body mass index.

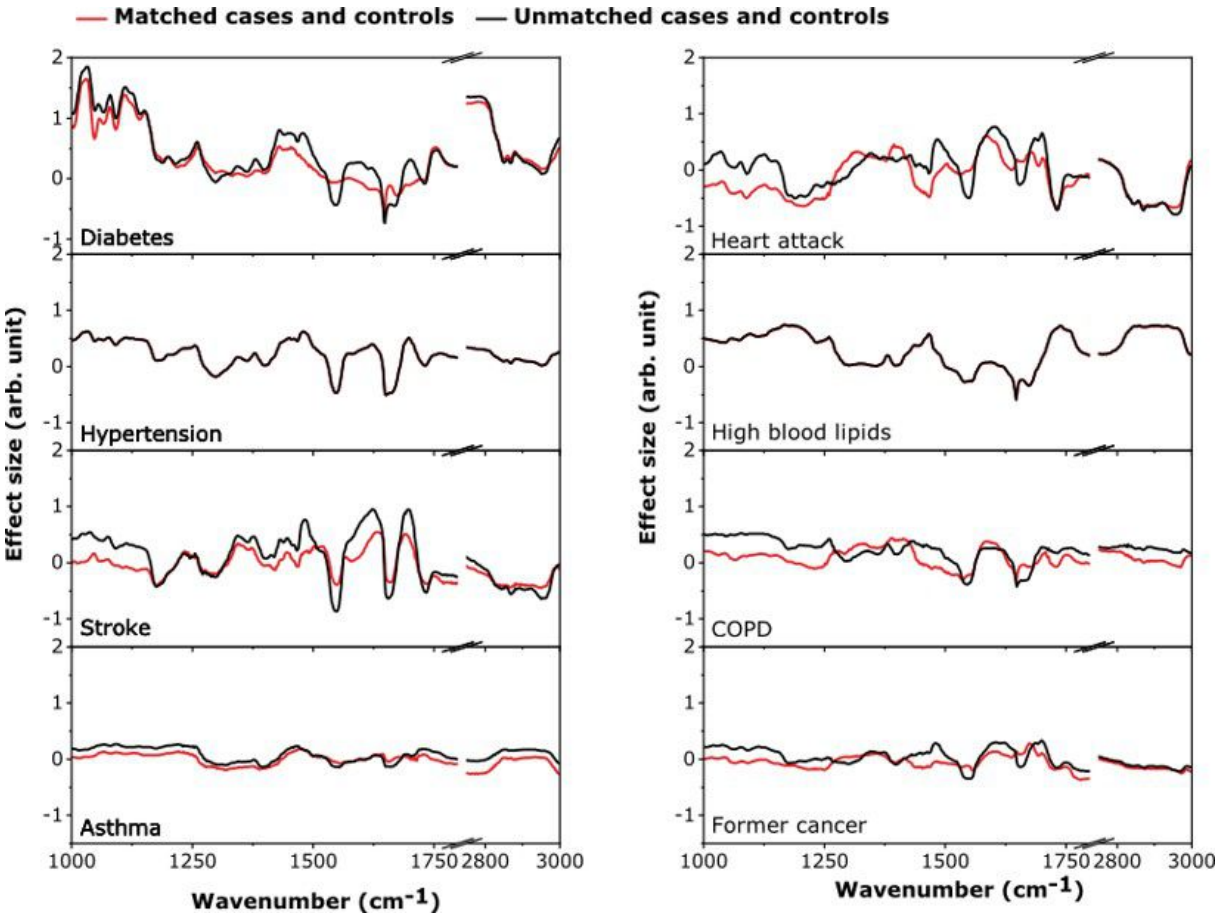

Figure 6.2: Effect size of common medical conditions on the FT-IR data of the KORA-FF4 cohort. The effect size obtained for unmatched cases and controls is always higher than for matched cases and controls, in agreement with what has been observed for the classification efficiencies and the SVM coefficients (section 4.2.3). For a high number of cases, such as for hypertension and high blood lipids, the differences are very small because, after matching, only a few controls are removed from the classification. The intensity of the effect size reflects the trend observed for the AUCs (Figure 4.6b). Acronyms: SVM - support vector machine; AUC - area under the curve; COPD - chronic obstructive pulmonary disease.

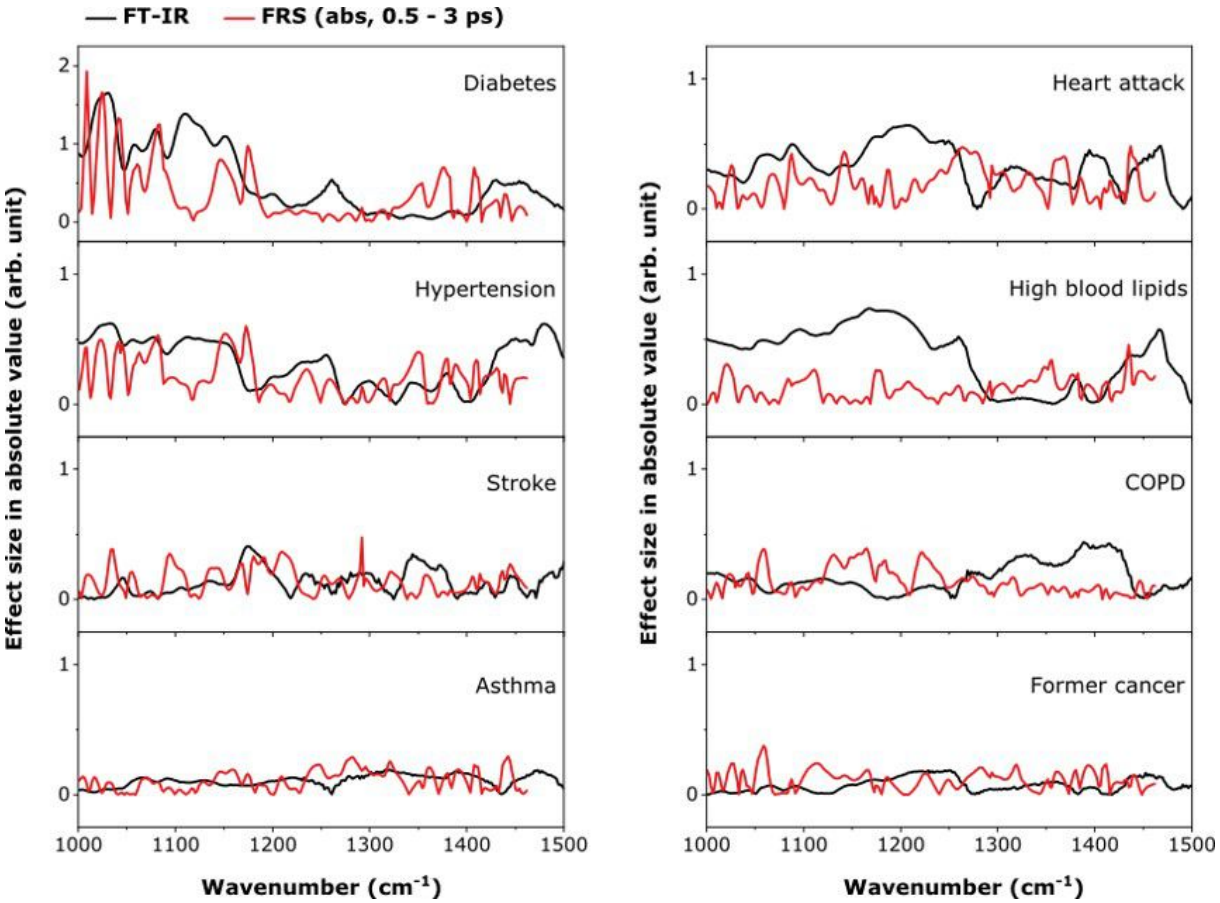

Figure 6.3: Effect size of common medical conditions on the FT-IR and FRS data of the KORA-FF4 cohort. The effect size is reported for the two techniques for matched cases and controls. For the FRS data, the effect size has been calculated on the "absorption" signal derived for the temporal window between 0.5 and 3 *ps*, where the AUCs are generally higher for all binary classifications. The discrepancies found for the two techniques can be attributed to the different nature of the signals, especially due to the temporal filter applied on the FRS data. Acronyms: FRS - field-resolved spectroscopy; AUC - area under the curve; COPD - chronic obstructive pulmonary disease.

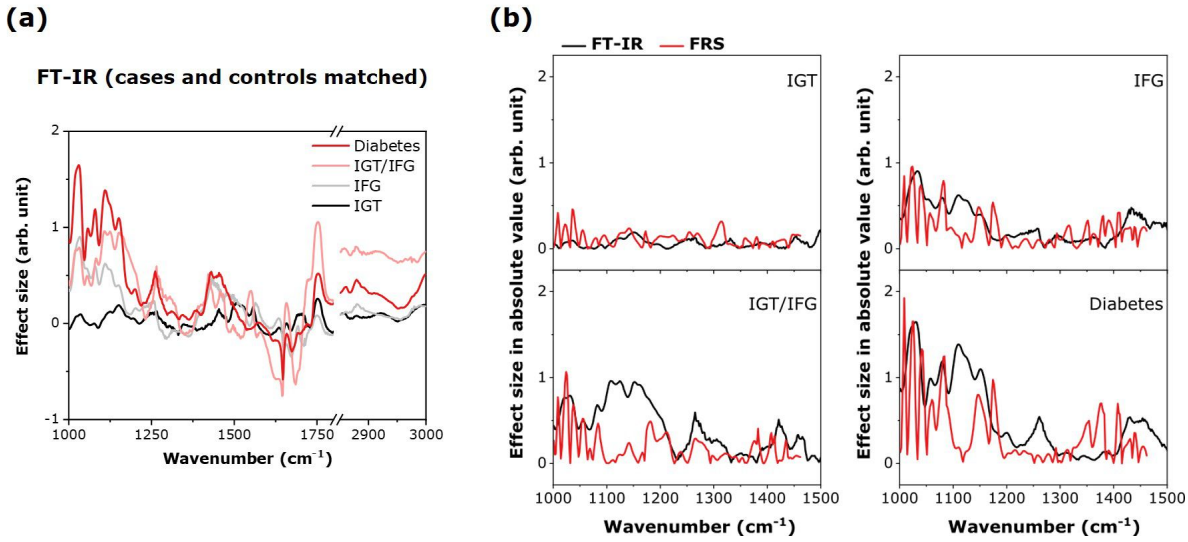

Figure 6.4: Effect size of diabetes and prediabetes on the FT-IR and FRS data of the KORA-FF4 cohort. (a) The effect size is reported for diabetes and three prediabetes entities with matched controls for the FT-IR data in the full spectral coverage (1000 - 3000 cm^{-1}). The intensity of the effect size is in agreement with the AUCs reported in the main text and highlights similar frequency patterns as the correspondent SVM coefficients (Figure 4.13b, c). (b) The effect size shown for FT-IR in panel (a) is compared, in the same spectral coverage ((1000 - 1500 cm^{-1}), with the effect size obtained for the FRS data (temporal filter: 0.5 - 3 ps). The man signatures are comparable between the FT-IR and FRS data, with discreoancies due to the different nature of the two signals, especially because of the temporal filter applied on the FRS data. The higher intensity of the effect size for FRS compared to FT-IR for the prediabetes condition of IGT is in agreement with what observed in the main text (Figure 4.15a). Acronyms: FRS - field-resolved spectroscopy; SVM - support vector machine; AUC - area under the curve; IGT - impaired glucose tolerance; IFG - impaired fasting glucose.

List of Publications

1. Meneghin, E., **Leonardo, C.**, Volpato, A., Bolzonello, L., Collini, E. (2017) "Mechanistic insight into internal conversion process within Q-bands of chlorophyll a." *Scientific reports*, 7(1), 11389

2. Voronina, L., **Leonardo, C.**, Mueller-Reif, J. B., Geyer, P. E., Huber, M., Trubetskov, M., Kepesidis, K. V., Behr, J., Mann, M., Krausz, F., Žigman, M. (2021) "Molecular Origin of Blood-based Infrared Spectroscopic Fingerprints." *Angewandte Chemie (International ed. in English)* 10.1002/anie.202103272. Advance online publication.

References

[1] H. Randall, *"Infra-red spectroscopy"*, Science **65**, 167–173 (1927).

[2] *"The infracord double beam spectrophotometer"*, Chemical & Engineering News Archive **35**, 74 (1957) 10.1021/cen-v035n033.p074.

[3] N. K. FREEMAN, "Infrared spectrometry", in, Vol. 4, edited by J. H. LAWRENCE and C. A. TOBIAS, Advances in Biological and Medical Physics (Elsevier, 1956), pp. 167–221, https://doi.org/10.1016/B978-1-4832-3110-5.50009-1.

[4] J. Clausen, H. Dyggve, and J. Melchior, *"Mucopolysaccharidosis: paper electrophoretic and infra-red analysis of the urine in gargoylism and morquio-ullrich's disease"*, Archives of disease in childhood **38**, 364 (1963).

[5] P. R. Griffiths and J. A. De Haseth, *Fourier transform infrared spectrometry*, Vol. 171 (John Wiley & Sons, 2007).

[6] NCBI, https://pubmed.ncbi.nlm.nih.gov/.

[7] WHO, *Global health and aging*, (2011) https://www.who.int/ageing/.

[8] V. Kulasingam and E. P. Diamandis, *"Strategies for discovering novel cancer biomarkers through utilization of emerging technologies"*, Nature clinical practice Oncology **5**, 588–599 (2008).

[9] I. Belczacka, A. Latosinska, J. Metzger, D. Marx, A. Vlahou, H. Mischak, and M. Frantzi, *"Proteomics biomarkers for solid tumors: current status and future prospects"*, Mass spectrometry reviews **38**, 49–78 (2019).

[10] A.-h. Zhang, H. Sun, S. Qiu, and X.-j. Wang, *"Metabolomics in noninvasive breast cancer"*, Clinica Chimica Acta **424**, 3–7 (2013).

[11] Y. Qiu, B. Zhou, M. Su, S. Baxter, X. Zheng, X. Zhao, Y. Yen, and W. Jia, *"Mass spectrometry-based quantitative metabolomics revealed a distinct lipid profile in breast cancer patients"*, International journal of molecular sciences **14**, 8047–8061 (2013).

[12] E. Heitzer and M. R. Speicher, *"One size does not fit all: size-based plasma dna diagnostics"*, Science Translational Medicine **10** (2018).

[13] M. Murph, T. Tanaka, J. Pang, E. Felix, S. Liu, R. Trost, A. K. Godwin, R. Newman, and G. Mills, *"Liquid chromatography mass spectrometry for quantifying plasma lysophospholipids: potential biomarkers for cancer diagnosis"*, Methods in enzymology **433**, 1–25 (2007).

[14] A. A. Bunaciu, H. Y. Aboul-Enein, and Ş. Fleschin, *"Vibrational spectroscopy in clinical analysis"*, Applied Spectroscopy Reviews **50**, 176–191 (2015).

[15] M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. A. Lukaszewski, G. Thiéfin, and G. D. Sockalingum, *"Developing and understanding biofluid vibrational spectroscopy: a critical review"*, Chemical Society Reviews **45**, 1803–1818 (2016).

[16] D. Finlayson, C. Rinaldi, and M. J. Baker, *"Is infrared spectroscopy ready for the clinic?"*, Analytical chemistry **91**, 12117–12128 (2019).

[17] F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad, and N. Nikbakhsh, *"Application of ft-ir spectroscopy on breast cancer serum analysis"*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **187**, 87–91 (2017).

[18] E. Barlev, U. Zelig, O. Bar, C. Segev, S. Mordechai, J. Kapelushnik, I. Nathan, F. Flomen, H. Kashtan, R. Dickman, et al., *"A novel method for screening colorectal cancer by infrared spectroscopy of peripheral blood mononuclear cells and plasma"*, Journal of gastroenterology **51**, 214–221 (2016).

[19] Y. Shen, A. Davies, E. Linfield, P. Taday, D. Arnone, and T. Elsey, *"Determination of glucose concentration in whole blood using ftir spectroscopy"*, Tera-Hertz Radiation in Biological Research, Investigation on Diagnostics, and Study on Potential Genotoxic Effects (2002).

[20] P. D. Lewis, K. E. Lewis, R. Ghosal, S. Bayliss, A. J. Lloyd, J. Wills, R. Godfrey, P. Kloer, and L. A. Mur, *"Evaluation of ftir spectroscopy as a diagnostic tool for lung cancer using sputum"*, BMC cancer **10**, 1–10 (2010).

[21] V. Untereiner, G. Dhruvananda Sockalingum, R. Garnotel, C. Gobinet, F. Ramaholimihaso, F. Ehrhard, M.-D. Diebold, and G. Thiéfin, *"Bile analysis using high-throughput ftir spectroscopy for the diagnosis of malignant biliary strictures: a pilot study in 57 patients"*, Journal of biophotonics **7**, 241–253 (2014).

[22] K.-Z. Liu, T. C. Dembinski, and H. H. Mantsch, *"Rapid determination of fetal lung maturity from infrared spectra of amniotic fluid"*, American journal of obstetrics and gynecology **178**, 234–241 (1998).

[23] D. Yonar, L. Ocek, B. I. Tiftikcioglu, Y. Zorlu, and F. Severcan, *"Relapsing-remitting multiple sclerosis diagnosis from cerebrospinal fluids via fourier transform infrared spectroscopy coupled with multivariate analysis"*, Scientific reports **8**, 1–13 (2018).

[24] K. V. Oliver, A. Vilasi, A. Maréchal, S. H. Moochhala, R. J. Unwin, and P. R. Rich, *"Infrared vibrational spectroscopy: a rapid and novel diagnostic and monitoring tool for cystinuria"*, Scientific reports **6**, 1–7 (2016).

[25] D. A. Scott, D. E. Renaud, S. Krishnasamy, P. Meriç, N. Buduneli, Ş. Çetinkalp, and K.-Z. Liu, *"Diabetes-related molecular signatures in infrared spectra of human saliva"*, Diabetology & metabolic syndrome **2**, 1–9 (2010).

[26] A. Travo, C. Paya, G. Déléris, J. Colin, B. Mortemousque, and I. Forfar, *"Potential of ftir spectroscopy for analysis of tears for diagnosis purposes"*, Analytical and bioanalytical chemistry **406**, 2367–2376 (2014).

[27] A. Khoshmanesh, M. W. Dixon, S. Kenny, L. Tilley, D. McNaughton, and B. R. Wood, *"Detection and quantification of early-stage malaria parasites in laboratory infected erythrocytes by attenuated total reflectance infrared spectroscopy and multivariate analysis"*, Analytical chemistry **86**, 4379–4386 (2014).

[28] A. Ergin, F. Großerüschkamp, O. Theisen, K. Gerwert, S. Remiszewski, C. M. Thompson, and M. Diem, *"A method for the comparison of multi-platform spectral histopathology (shp) data sets"*, Analyst **140**, 2465–2472 (2015).

[29] M. Diem, A. Ergin, S. Remiszewski, and X. Mu, *"Ps01. 31: a reagent-free, high resolution lung cancer diagnostic method based on phenotypic infrared spectral imaging: topic: pathology"*, Journal of Thoracic Oncology **11**, S288 (2016).

[30] *Glyconics - shaping the future of medical diagnostics*, https://glyconics.com/.

[31] S. Whiteman, Y. Yang, J. Jones, and M. Spiteri, *"Ftir spectroscopic analysis of sputum: preliminary findings on a potential novel diagnostic marker for copd"*, Therapeutic advances in respiratory disease **2**, 23–31 (2008).

[32] N. Patel, P. Coleborn, A. Hampton, V. Campbell, M. Allen, A. Gahkani, and M. Spiteri, *"P111 ftir spectroscopic profiling of copd sputum: identification of distinct spectral signatures and correlation to copd status"*, Thorax **65**, A124–A125 (2010).

[33] E. Gray, H. J. Butler, R. Board, P. M. Brennan, A. J. Chalmers, T. Dawson, J. Goodden, W. Hamilton, M. G. Hegarty, A. James, et al., *"Health economic evaluation of a serum-based blood test for brain tumour diagnosis: exploration of two clinical scenarios"*, BMJ open **8** (2018).

[34] C. Mann, *"Observational research methods. research design ii: cohort, cross sectional, and case-control studies"*, Emergency medicine journal **20**, 54–60 (2003).

[35] P. Sedgwick, *"Bias in observational study designs: cross sectional studies"*, Bmj **350** (2015).

[36] U. S. Kesmodel, *"Cross-sectional studies–what are they good for?"*, Acta obstetricia et gynecologica Scandinavica **97**, 388–393 (2018).

[37] J. I. Hudson, H. G. Pope Jr, and R. J. Glynn, *"The cross-sectional cohort study: an underutilized design"*, Epidemiology **16**, 355–359 (2005).

[38] I. Pupeza, D. Sánchez, J. Zhang, N. Lilienfein, M. Seidel, N. Karpowicz, T. Paasch-Colberg, I. Znakovskaya, M. Pescher, W. Schweinberger, et al., *"High-power sub-two-cycle mid-infrared pulses at 100 mhz repetition rate"*, Nature Photonics **9**, 721–724 (2015).

[39] C. Gaida, M. Gebhardt, T. Heuermann, F. Stutzki, C. Jauregui, J. Antonio-Lopez, A. Schülzgen, R. Amezcua-Correa, A. Tünnermann, I. Pupeza, et al., *"Watt-scale super-octave mid-infrared intrapulse difference frequency generation"*, Light: Science & Applications **7**, 1–8 (2018).

[40] M. Seidel, X. Xiao, S. A. Hussain, G. Arisholm, A. Hartung, K. T. Zawilski, P. G. Schunemann, F. Habel, M. Trubetskov, V. Pervak, et al., *"Multi-watt, multi-octave, mid-infrared femtosecond source"*, Science advances **4**, eaaq1526 (2018).

[41] T. Butler, D. Gerz, C. Hofer, J. Xu, C. Gaida, T. Heuermann, M. Gebhardt, L. Vamos, W. Schweinberger, J. Gessner, et al., *"Watt-scale 50-mhz source of single-cycle waveform-stable pulses in the molecular fingerprint region"*, Optics letters **44**, 1730–1733 (2019).

[42] K. Araki, N. Yagi, Y. Ikemoto, H. Yagi, C.-J. Choong, H. Hayakawa, G. Beck, H. Sumi, H. Fujimura, T. Moriwaki, et al., *"Synchrotron ftir micro-spectroscopy for structural analysis of lewy bodies in the brain of parkinson's disease patients"*, Scientific reports **5**, 1–8 (2015).

[43] E. Aboualizadeh, M. Ranji, C. M. Sorenson, R. Sepehr, N. Sheibani, and C. J. Hirschmugl, *"Retinal oxidative stress at the onset of diabetes determined by synchrotron ftir widefield imaging: towards diabetes pathogenesis"*, Analyst **142**, 1061–1072 (2017).

[44] M. Grzelak, P. Wróbel, M. Lankosz, Z. Stęgowski, D. Adamek, B. Hesse, H. Castillo-Michel, et al., *"Diagnosis of ovarian tumour tissues by sr-ftir spectroscopy: a pilot study"*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **203**, 48–55 (2018).

[45] F. Lyng, E. Gazi, and P. Gardner, *"Preparation of tissues and cells for infrared and raman spectroscopy and imaging"*, Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach Royal Society of Chemistry, Cambridge, 147–191 (2010).

[46] L. Buriankova, Z. Nadova, D. Jancura, M. Refregiers, I. Yousef, J. Mikes, and P. Miskovsky, *"Synchrotron based fourier-transform infrared microspectroscopy as sensitive technique for the detection of early apoptosis in u-87 mg cells"*, Laser Physics Letters **7**, 613 (2010).

[47] P. Werle, R. Mücke, and F. Slemr, *"The limits of signal averaging in atmospheric trace-gas monitoring by tunable diode-laser absorption spectroscopy (tdlas)"*, Applied Physics B **57**, 131–139 (1993).

[48] N. R. Newbury, I. Coddington, and W. Swann, *"Sensitivity of coherent dual-comb spectroscopy"*, Optics express **18**, 7929–7945 (2010).

[49] P. C. Hobbs, *"Ultrasensitive laser measurements without tears"*, Applied optics **36**, 903–920 (1997).

[50] M. Huber, W. Schweinberger, F. Stutzki, J. Limpert, I. Pupeza, and O. Pronin, *"Active intensity noise suppression for a broadband mid-infrared laser source"*, Optics express **25**, 22499–22509 (2017).

[51] J. Li, B. Yu, W. Zhao, and W. Chen, *"A review of signal enhancement and noise reduction techniques for tunable diode laser absorption spectroscopy"*, Applied Spectroscopy Reviews **49**, 666–691 (2014).

[52] I. Pupeza, M. Huber, M. Trubetskov, W. Schweinberger, S. A. Hussain, C. Hofer, K. Fritsch, M. Poetzlberger, L. Vamos, E. Fill, et al., *"Field-resolved infrared spectroscopy of biological systems"*, Nature **577**, 52–59 (2020).

[53] C. Meisinger, A. Peters, and J. Linseisen, *"Vom monica-projekt ueber kora zur nakostudie: vom praktischen nutzen von bevoelkerungsstudien in der region augsburg"*, Das Gesundheitswesen **78**, 84–90 (2016).

[54] R. Holle, M. Happich, H. Loewel, H.-E. Wichmann, null for the MONICA/KORA Study Group, et al., *"Kora-a research platform for population based health research"*, Das Gesundheitswesen **67**, 19–25 (2005).

[55] W. Hoffmann, U. Latza, C. Terschüren, et al., *"Guidelines and recommendations for ensuring good epidemiological practice (gep)–revised version after evaluation"*, Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)) **67**, 217 (2005).

[56] R. Holle, B. Giesecke, and H. Nagl, *"Pc-gestuetzte datenerhebung als beitrag zur qualitaetssicherung in gesundheitssurveys: erfahrungen mit daimon im kora-survey 2000"*, Zeitschrift für Gesundheitswissenschaften= Journal of public health **8**, 165–173 (2000).

[57] N. Muehlberger, C. Behrend, R. Stark, and R. Holle, *"Datenbankgestuetzte online-erfassung von arzneimitteln im rahmen gesundheitswissenschaftlicher studien-erfahrungen mit der idom-software"*, Inform Biom Epidemiol Med Biol **34**, 601–611 (2003).

[58] L. Lovergne, P. Bouzy, V. Untereiner, R. Garnotel, M. Baker, G. Thiéfin, and G. Sockalingum, *"Biofluid infrared spectro-diagnostics: pre-analytical considerations for clinical applications"*, Faraday discussions **187**, 521–537 (2016).

[59] M. Plebani and P. Carraro, *"Mistakes in a stat laboratory: types and frequency"*, Clinical chemistry **43**, 1348–1351 (1997).

[60] M. e. a. Huber, *"Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring, submitted"*, Nature communications (2021).

[61] M. Diem, *"Comments on recent reports on infrared spectral detection of disease markers in blood components"*, Journal of Biophotonics **11**, e201800064 (2018).

[62] L. Voronina, C. Leonardo, J. B. Mueller-Reif, P. E. Geyer, M. Huber, M. Trubetskov, K. V. Kepesidis, J. Behr, M. Mann, F. Krausz, et al., *"Molecular origin of blood-based infrared spectroscopic fingerprints"*, Angewandte Chemie International Edition (2021).

[63] J. d. Paula, *Atkins'physical chemistry*, 2006.

[64] W. W. Parson, *Modern optical spectroscopy*, Vol. 2 (Springer, 2007).

[65] P. M. Morse, *"Diatomic molecules according to the wave mechanics. ii. vibrational levels"*, Physical review **34**, 57 (1929).

[66] N. Colthup, *Introduction to infrared and raman spectroscopy* (Elsevier, 2012).

[67] P. Connes, *"Early history of fourier transform spectroscopy"*, Infrared Physics **24**, 69–93 (1984).

[68] M. Diem, *Modern vibrational spectroscopy and micro-spectroscopy: theory, instrumentation and biomedical applications* (John Wiley & Sons, 2015).

[69] "Surface-enhanced infrared spectroscopy", in *Encyclopedia of biophysics*, edited by G. C. K. Roberts (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp. 2536–2536, 10.1007/978-3-642-16712-6_101019.

[70] C. G. Atkins, K. Buckley, M. W. Blades, and R. F. Turner, *"Raman spectroscopy of blood and blood components"*, Applied spectroscopy **71**, 767–793 (2017).

[71] M. J. Baker, H. J. Byrne, J. Chalmers, P. Gardner, R. Goodacre, A. Henderson, S. G. Kazarian, F. L. Martin, J. Moger, N. Stone, et al., *"Clinical applications of infrared and raman spectroscopy: state of play and future challenges"*, Analyst **143**, 1735–1757 (2018).

[72] A. Schwaighofer, M. Brandstetter, and B. Lendl, *"Quantum cascade lasers (qcls) in biomedical spectroscopy"*, Chemical Society Reviews **46**, 5903–5924 (2017).

[73] M. Huber, M. Trubetskov, S. A. Hussain, W. Schweinberger, C. Hofer, and I. Pupeza, *"Optimum sample thickness for trace analyte detection with field-resolved infrared spectroscopy"*, Analytical chemistry **92**, 7508–7514 (2020).

[74] G. van Rossum, *"Python tutorial, technical report cs-r9526, centrum voor wiskunde en informatica (cwi), amsterdam."*, (1995).

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *"Scikit-learn: machine learning in python"*, the Journal of machine Learning research **12**, 2825–2830 (2011).

[76] R. Team et al., *"Rstudio: integrated development for r"*, RStudio, Inc., Boston, MA URL http://www. rstudio. com **42**, 84 (2015).

[77] S. Y. Kung, *Kernel methods and machine learning* (Cambridge University Press, 2014), 10.1017/CBO9781139176224.

[78] D. Forsyth, *Applied machine learning* (Springer, 2019).

[79] S. Y. Kung, *Kernel methods and machine learning* (Cambridge University Press, 2014).

[80] I. Guyon and A. Elisseeff, *"An introduction to variable and feature selection"*, Journal of machine learning research **3**, 1157–1182 (2003).

[81] P. R. Rosenbaum et al., *Design of observational studies*, Vol. 10 (Springer, 2010).

[82] A. Olmos and P. Govindasamy, *"Propensity scores: a practical introduction using r"*, Journal of MultiDisciplinary Evaluation **11**, 68–88 (2015).

[83] G. Anton, R. Wilson, Z.-h. Yu, C. Prehn, S. Zukunft, J. Adamski, M. Heier, C. Meisinger, W. Römisch-Margl, R. Wang-Sattler, et al., *"Pre-analytical sample quality: metabolite ratios as an intrinsic marker for prolonged room temperature exposure of serum samples"*, PloS one **10**, e0121495 (2015).

[84] L. Lovergne, J. Lovergne, P. Bouzy, V. Untereiner, M. Offroy, R. Garnotel, G. Thiéfin, M. J. Baker, and G. D. Sockalingum, *"Investigating pre-analytical requirements for serum and plasma based infrared spectro-diagnostic"*, Journal of Biophotonics **12**, e201900177 (2019).

[85] S. F. Green, *"The cost of poor blood specimen quality and errors in preanalytical processes"*, Clinical biochemistry **46**, 1175–1179 (2013).

[86] T. Sangster, H. Major, R. Plumb, A. J. Wilson, and I. D. Wilson, *"A pragmatic and readily implemented quality control strategy for hplc-ms and gc-ms-based metabonomic analysis"*, Analyst **131**, 1075–1078 (2006).

[87] H. Yang, S. Yang, J. Kong, A. Dong, and S. Yu, *"Obtaining information about protein secondary structures in aqueous solution using fourier transform ir spectroscopy"*, Nature protocols **10**, 382–396 (2015).

[88] D. J. Segelstein, "The complex refractive index of water", PhD thesis (University of Missouri–Kansas City, 1981).

[89] N. R. Sproston and J. J. Ashworth, *"Role of c-reactive protein at sites of inflammation and infection"*, Frontiers in immunology **9**, 754 (2018).

[90] C. Petibois, G. Déléris, and G. Cazorla, *"Perspectives in the utilisation of fourier-transform infrared spectroscopy of serum in sports medicine"*, Sports Medicine **29**, 387–396 (2000).

[91] C. Petibois, G. Cazorla, and G. Déléris, *"Ft-ir spectroscopy utilization to sportsmen fatigability evaluation and control."*, Medicine and science in sports and exercise **32**, 1803 (2000).

[92] B. Suh, S. Park, D. W. Shin, J. M. Yun, B. Keam, H.-K. Yang, E. Ahn, H. Lee, J. Park, and B. Cho, *"Low albumin-to-globulin ratio associated with cancer incidence and mortality in generally healthy adults"*, Annals of oncology **25**, 2260–2266 (2014).

[93] A. A. Johnson and A. Stolzing, *"The role of lipid metabolism in aging, lifespan regulation, and age-related disease"*, Aging Cell **18**, e13048 (2019).

[94] I. Gomi, H. Fukushima, M. Shiraki, Y. Miwa, T. Ando, K. Takai, and H. Moriwaki, *"Relationship between serum albumin level and aging in community-dwelling self-supported elderly population"*, Journal of nutritional science and vitaminology **53**, 37–42 (2007).

[95] J. S. Yudkin, C. Stehouwer, J. Emeis, and S. Coppack, *"C-reactive protein in healthy subjects: associations with obesity, insulin resistance, and endothelial dysfunction: a potential role for cytokines originating from adipose tissue?"*, Arteriosclerosis, thrombosis, and vascular biology **19**, 972–978 (1999).

[96] E. Selvin, N. P. Paynter, and T. P. Erlinger, *"The effect of weight loss on c-reactive protein: a systematic review"*, Archives of internal medicine **167**, 31–39 (2007).

[97] S. Ahadi, W. Zhou, S. M. S.-F. Rose, M. R. Sailani, K. Contrepois, M. Avina, M. Ashland, A. Brunet, and M. Snyder, *"Personal aging markers and ageotypes revealed by deep longitudinal profiling"*, Nature medicine **26**, 83–90 (2020).

[98] C. Franceschi and J. Campisi, *"Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases"*, Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences **69**, S4–S9 (2014).

[99] C. M. Chak, M. E. Lacruz, J. Adam, S. Brandmaier, M. Covic, J. Huang, C. Meisinger, D. Tiller, C. Prehn, J. Adamski, et al., *"Ageing investigation using two-time-point metabolomics data from kora and carla studies"*, Metabolites **9**, 44 (2019).

[100] G. Kolovou, N. Katsiki, A. Pavlidis, H. Bilianou, G. Goumas, and D. P Mikhailidis, *"Ageing mechanisms and associated lipid changes"*, Current vascular pharmacology **12**, 682–689 (2014).

[101] G. Alter, T. H. Ottenhoff, and S. A. Joosten, "Antibody glycosylation in inflammation, disease and vaccination", in Seminars in immunology, Vol. 39 (Elsevier, 2018), pp. 102–110.

[102] C. Reily, T. J. Stewart, M. B. Renfrow, and J. Novak, *"Glycosylation in health and disease"*, Nature Reviews Nephrology **15**, 346–366 (2019).

[103] I. Gudelj, G. Lauc, and M. Pezer, *"Immunoglobulin g glycosylation in aging and diseases"*, Cellular immunology **333**, 65–79 (2018).

[104] Y. Miura and T. Endo, *"Glycomics and glycoproteomics focused on aging and age-related diseases—glycans as a potential biomarker for physiological alterations"*, Biochimica et Biophysica Acta (BBA)-General Subjects **1860**, 1608–1614 (2016).

[105] A. V. Everest-Dass, E. S. Moh, C. Ashwood, A. M. Shathili, and N. H. Packer, *"Human disease glycomics: technology advances enabling protein glycosylation analysis–part 2"*, Expert review of proteomics **15**, 341–352 (2018).

[106] Z. Chan, Y. C. Chooi, C. Ding, J. Choo, S. A. Sadananthan, N. Michael, S. S. Velan, M. K. Leow, and F. Magkos, *"Sex differences in glucose and fatty acid metabolism in asians who are nonobese"*, The Journal of Clinical Endocrinology & Metabolism **104**, 127–136 (2019).

[107] K. A. Lawton, A. Berger, M. Mitchell, K. E. Milgram, A. M. Evans, L. Guo, R. W. Hanson, S. C. Kalhan, J. A. Ryals, and M. V. Milburn, *"Analysis of the adult human plasma metabolome"*, (2008).

[108] A. Bardelli and K. Pantel, *"Liquid biopsies, what we do not know (yet)"*, Cancer cell **31**, 172–179 (2017).

[109] P. Lasch and J. Kneipp, *Biomedical vibrational spectroscopy* (John Wiley & Sons, 2008).

[110] M. Paraskevaidi, C. L. Morais, K. M. Lima, J. S. Snowden, J. A. Saxon, A. M. Richardson, M. Jones, D. M. Mann, D. Allsop, P. L. Martin-Hirsch, et al., *"Differential diagnosis of alzheimer's disease using spectrochemical analysis of blood"*, Proceedings of the National Academy of Sciences **114**, E7929–E7938 (2017).

[111] K. Thumanu, S. Sangrajrang, T. Khuhaprema, A. Kalalak, W. Tanthanuch, S. Pongpiachan, and P. Heraud, *"Diagnosis of liver cancer from blood sera using ftir microspectroscopy: a preliminary study"*, Journal of biophotonics **7**, 222–231 (2014).

[112] I. Taleb, G. Thiéfin, C. Gobinet, V. Untereiner, B. Bernard-Chabert, A. Heurgué, C. Truntzer, P. Hillon, M. Manfait, P. Ducoroy, et al., *"Diagnosis of hepatocellular carcinoma in cirrhotic patients: a proof-of-concept study using serum micro-raman spectroscopy"*, Analyst **138**, 4006–4014 (2013).

[113] U. Zelig, E. Barlev, O. Bar, I. Gross, F. Flomen, S. Mordechai, J. Kapelushnik, I. Nathan, H. Kashtan, N. Wasserberg, et al., *"Early detection of breast cancer using total biochemical analysis of peripheral blood components: a preliminary study"*, BMC cancer **15**, 1–10 (2015).

[114] K. M. Lima, K. B. Gajjar, P. L. Martin-Hirsch, and F. L. Martin, *"Segregation of ovarian cancer stage exploiting spectral biomarkers derived from blood plasma or serum analysis: atr-ftir spectroscopy coupled with variable selection methods"*, Biotechnology progress **31**, 832–839 (2015).

[115] J. R. Hands, G. Clemens, R. Stables, K. Ashton, A. Brodbelt, C. Davis, T. P. Dawson, M. D. Jenkinson, R. W. Lea, C. Walker, et al., *"Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection fourier-transform infrared spectroscopy"*, Journal of neuro-oncology **127**, 463–472 (2016).

[116] P. Carmona, M. Molina, M. Calero, F. Bermejo-Pareja, P. Martinez-Martin, and A. Toledano, *"Discrimination analysis of blood plasma associated with alzheimer's disease using vibrational spectroscopy"*, Journal of Alzheimer's Disease **34**, 911–920 (2013).

[117] J. Ollesch, D. Theegarten, M. Altmayer, K. Darwiche, T. Hager, G. Stamatis, and K. Gerwert, *"An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma"*, Biomedical Spectroscopy and Imaging **5**, 129–144 (2016).

[118] X. Yang, T. Fang, Y. Li, L. Guo, F. Li, F. Huang, and L. Li, *"Pre-diabetes diagnosis based on atr-ftir spectroscopy combined with cart and xgboots"*, Optik **180**, 189–198 (2019).

[119] S. L. Haas, R. Müller, A. Fernandes, K. Dzeyk-Boycheva, S. Würl, J. Hohmann, S. Hemberger, E. Elmas, M. Brückmann, P. Bugert, et al., *"Spectroscopic diagnosis of myocardial infarction and heart failure by fourier transform infrared spectroscopy in serum samples"*, Applied spectroscopy **64**, 262–267 (2010).

[120] M. Baker, *"Photonic biofluid diagnostics"*, Journal of Biophotonics **7**, 151–152 (2014).

[121] C. G. Fraser, *"Inherent biological variation and reference values"*, Clinical Chemistry and Laboratory Medicine (CCLM) **42**, 758–764 (2004).

[122] I. S. Ockene, D. E. Chiriboga, E. J. Stanek III, M. G. Harmatz, R. Nicolosi, G. Saperia, A. D. Well, P. Freedson, P. A. Merriam, G. Reed, et al., *"Seasonal variation in serum cholesterol levels: treatment implications and possible mechanisms"*, Archives of internal medicine **164**, 863–870 (2004).

[123] J. N. Sampson, S. M. Boca, X. O. Shu, R. Z. Stolzenberg-Solomon, C. E. Matthews, A. W. Hsing, Y. T. Tan, B.-T. Ji, W.-H. Chow, Q. Cai, et al., *"Metabolomics in epidemiology: sources of variability in metabolite measurements and implications"*, Cancer Epidemiology and Prevention Biomarkers **22**, 631–640 (2013).

[124] J. M. Ramsey, J. D. Cooper, B. W. Penninx, and S. Bahn, *"Variation in serum biomarkers with sex and female hormonal status: implications for clinical tests"*, Scientific reports **6**, 26947 (2016).

[125] S. De Bruyne, T. Monteyne, M. M. Speeckaert, and J. R. Delanghe, *"Infrared analysis of lipoproteins in the detection of alcohol biomarkers"*, Clinical Chemistry and Laboratory Medicine (CCLM) **55**, 876–881 (2017).

[126] K. Sharma, S. P. Sharma, and S. C. Lahiri, *"Estimation of blood alcohol concentration by horizontal attenuated total reflectance–fourier transform infrared spectroscopy"*, Alcohol **44**, 351–357 (2010).

[127] K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch, and F. L. Martin, *"Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer"*, Analyst **138**, 3917–3926 (2013).

[128] Z. Yu, G. Kastenmüller, Y. He, P. Belcredi, G. Möller, C. Prehn, J. Mendes, S. Wahl, W. Roemisch-Margl, U. Ceglarek, et al., *"Differences between human plasma and serum metabolite profiles"*, PloS one **6**, e21230 (2011).

[129] S. Yazdanpanah, M. Rabiee, M. Tahriri, M. Abdolrahim, A. Rajab, H. E. Jazayeri, and L. Tayebi, *"Evaluation of glycated albumin (ga) and ga/hba1c ratio for diagnosis of diabetes and glycemic control: a comprehensive review"*, Critical reviews in clinical laboratory sciences **54**, 219–232 (2017).

[130] S. P. Rauh, M. W. Heymans, A. D. Koopman, G. Nijpels, C. D. Stehouwer, B. Thorand, W. Rathmann, C. Meisinger, A. Peters, T. De Las Heras Gala, et al., *"Predicting glycated hemoglobin levels in the non-diabetic general population: development and validation of the direct-detect prediction model-a direct study"*, PLoS One **12**, e0171816 (2017).

[131] P. Guang, W. Huang, L. Guo, X. Yang, F. Huang, M. Yang, W. Wen, and L. Li, *"Blood-based ftir-atr spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: a stard compliant diagnosis research"*, Medicine **99** (2020).

[132] B. V. Howard, D. C. Robbins, M. L. Sievers, E. T. Lee, D. Rhoades, R. B. Devereux, L. D. Cowan, R. S. Gray, T. K. Welty, O. T. Go, et al., *"Ldl cholesterol as a strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low ldl: the strong heart study"*, Arteriosclerosis, thrombosis, and vascular biology **20**, 830–835 (2000).

[133] H. A. Khan, A. Ekhzaimy, I. Khan, and M. K. Sakharkar, *"Potential of lipoproteins as biomarkers in acute myocardial infarction"*, Anatolian journal of cardiology **18**, 68 (2017).

[134] K.-Z. Liu, R. A. Shaw, A. Man, T. C. Dembinski, and H. H. Mantsch, *"Reagent-free, simultaneous determination of serum cholesterol in hdl and ldl by infrared spectroscopy"*, Clinical chemistry **48**, 499–506 (2002).

[135] A. Shamshirian, R. Alizadeh-Navaei, S. Abedi, H. Jafarpour, H. Fazli, S. Hosseini, A. Hessami, K. Karimifar, S. Yosefi, M. Zahedi, et al., *"Levels of blood biomarkers among patients with myocardial infarction in comparison to control group"*, Ethiopian journal of health sciences **30** (2020).

[136] D. Pandya, A. K. Nagrajappa, and K. Ravi, *"Assessment and correlation of urea and creatinine levels in saliva and serum of patients with chronic kidney disease, diabetes and hypertension–a research study"*, Journal of clinical and diagnostic research: JCDR **10**, ZC58 (2016).

[137] J. LaRosa, D. Hunninghake, D. Bush, M. Criqui, G. Getz, A. Gotto Jr, S. M. Grundy, L. Rakita, R. Robertson, and M. Weisfeldt, *"The cholesterol facts. a summary of the evidence relating dietary fats, serum cholesterol, and coronary heart disease. a joint statement by the american heart association and the national heart, lung, and blood institute. the task force on cholesterol issues, american heart association."*, Circulation **81**, 1721–1733 (1990).

[138] H. Kolb and S. Martin, *"Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes"*, BMC medicine **15**, 1–11 (2017).

[139] L. J. Appel, *"The effects of dietary factors on blood pressure"*, Cardiology clinics **35**, 197–212 (2017).

[140] C. Petibois, G. Cazorla, and G. Déléris, *"Triglycerides and glycerol concentration determinations using plasma ft-ir spectra"*, Applied spectroscopy **56**, 10–16 (2002).

[141] A. Langsted, B. G. Nordestgaard, and P. R. Kamstrup, *"Elevated lipoprotein (a) and risk of ischemic stroke"*, Journal of the American College of Cardiology **74**, 54–66 (2019).

[142] V. A. Goulart, M. M. Sena, T. O. Mendes, H. C. Menezes, Z. L. Cardeal, M. J. Paiva, V. C. Sandrim, M. C. Pinto, and R. R. Resende, *"Amino acid biosignature in plasma among ischemic stroke subtypes"*, BioMed research international **2019** (2019).

[143] Y. Xiao, Y. Yuan, Y. Liu, Y. Yu, N. Jia, L. Zhou, H. Wang, S. Huang, Y. Zhang, H. Yang, et al., *"Circulating multiple metals and incident stroke in chinese adults: the dongfeng-tongji cohort"*, Stroke **50**, 1661–1668 (2019).

[144] D. Moher, C. S. Dulberg, and G. A. Wells, *"Statistical power, sample size, and their reporting in randomized controlled trials"*, Jama **272**, 122–124 (1994).

[145] K. Suresh and S. Chandrashekara, *"Sample size estimation and power analysis for clinical research studies"*, Journal of human reproductive sciences **5**, 7 (2012).

[146] Y. Huang, *"Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case–control studies"*, Biostatistics **17**, 499–522 (2016).

[147] A. K. Jain and B. Chandrasekaran, *"39 dimensionality and sample size considerations in pattern recognition practice"*, Handbook of statistics **2**, 835–855 (1982).

[148] S. Roudys and A. Jain, *"Samll sample size effect in statistical pattern recognition"*, IEEE-Transactions on Pattern Analysis and Machine Intelligence **13**, 252–264 (1991).

[149] H. Kalayeh and D. A. Landgrebe, *"Predicting the required number of training samples"*, IEEE transactions on pattern analysis and machine intelligence, 664–667 (1983).

[150] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, *"Sample size planning for classification models"*, Analytica chimica acta **760**, 25–33 (2013).

[151] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov, *"Estimating dataset size requirements for classifying dna microarray data"*, Journal of computational biology **10**, 119–142 (2003).

[152] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, *"Predicting sample size required for classification performance"*, BMC medical informatics and decision making **12**, 1–10 (2012).

[153] C. M. Edwards and K. Cusi, *"Prediabetes: a worldwide epidemic"*, Endocrinology and Metabolism Clinics **45**, 751–764 (2016).

[154] M. L. Wilson, *"Prediabetes: beyond the borderline"*, Nursing Clinics **52**, 665–677 (2017).

[155] Y. Huang, X. Cai, W. Mai, M. Li, and Y. Hu, *"Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis"*, Bmj **355**, i5953 (2016).

[156] A. Zand, K. Ibrahim, and B. Patham, *"Prediabetes: why should we care?"*, Methodist DeBakey cardiovascular journal **14**, 289 (2018).

[157] W. H. Organization et al., *"Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a who/idf consultation"*, (2006).

[158] K. Pippitt, M. Li, and H. E. Gurgle, *"Diabetes mellitus: screening and diagnosis"*, American family physician **93**, 103–109 (2016).

[159] R. Wang-Sattler, Z. Yu, C. Herder, A. C. Messias, A. Floegel, Y. He, K. Heim, M. Campillos, C. Holzapfel, B. Thorand, et al., *"Novel biomarkers for pre-diabetes identified by metabolomics"*, Molecular systems biology **8**, 615 (2012).

[160] A. J. Hanley, K. Williams, C. Gonzalez, R. B. D'Agostino, L. E. Wagenknecht, M. P. Stern, and S. M. Haffner, *"Prediction of type 2 diabetes using simple measures of insulin resistance: combined results from the san antonio heart study, the mexico city diabetes study, and the insulin resistance atherosclerosis study"*, Diabetes **52**, 463–469 (2003).

[161] A. L. Bortheiry, D. A. Malerbi, and L. J. Franco, *"The roc curve in the evaluation of fasting capillary blood glucose as a screening test for diabetes and igt"*, Diabetes Care **17**, 1269–1272 (1994).

[162] Y. Hu, W. Liu, Y. Chen, M. Zhang, L. Wang, H. Zhou, P. Wu, X. Teng, Y. Dong, J. wen Zhou, et al., *"Combined use of fasting plasma glucose and glycated hemoglobin a1c in the screening of diabetes and impaired glucose tolerance"*, Acta diabetologica **47**, 231–236 (2010).

[163] C. Weyer, R. L. Hanson, P. A. Tataranni, C. Bogardus, and R. E. Pratley, *"A high fasting plasma insulin concentration predicts type 2 diabetes independent of insulin resistance: evidence for a pathogenic role of relative hyperinsulinemia."*, Diabetes **49**, 2094–2101 (2000).

[164] C. Wildberg, A. Masuch, K. Budde, G. Kastenmüller, A. Artati, W. Rathmann, J. Adamski, T. Kocher, H. Völzke, M. Nauck, et al., *"Plasma metabolomics to identify and stratify patients with impaired glucose tolerance"*, The Journal of Clinical Endocrinology & Metabolism **104**, 6357–6370 (2019).

[165] J. Long, L. Liu, Q. Jia, Z. Yang, Z. Sun, C. Yan, and D. Yan, *"Integrated biomarker for type 2 diabetes mellitus and impaired fasting glucose based on metabolomics analysis using ultra-high performance liquid chromatography quadrupole-orbitrap high-resolution accurate mass spectrometry"*, Rapid Communications in Mass Spectrometry **34**, e8779 (2020).

[166] D. Bogdanet, P. O'Shea, C. Lyons, A. Shafat, and F. Dunne, *"The oral glucose tolerance test—is it time for a change?—a literature review with an emphasis on pregnancy"*, Journal of Clinical Medicine **9**, 3451 (2020).

[167] M. J. McNeely, E. J. Boyko, D. L. Leonetti, S. E. Kahn, and W. Y. Fujimoto, *"Comparison of a clinical model, the oral glucose tolerance test, and fasting glucose for prediction of type 2 diabetes risk in japanese americans"*, Diabetes Care **26**, 758–763 (2003).

[168] M. Geva, G. Shlomai, A. Berkovich, E. Maor, A. Leibowitz, A. Tenenbaum, and E. Grossman, *"The association between fasting plasma glucose and glycated hemoglobin in the prediabetes range and future development of hypertension"*, Cardiovascular diabetology **18**, 1–9 (2019).

[169] Y. Shen, A. Davies, E. Linfield, T. Elsey, P. Taday, and D. Arnone, *"The use of fourier-transform infrared spectroscopy for the quantitative determination of glucose concentration in whole blood"*, Physics in Medicine & Biology **48**, 2023 (2003).

[170] S. Rassel, C. Xu, S. Zhang, and D. Ban, *"Noninvasive blood glucose detection using a quantum cascade laser"*, Analyst **145**, 2441–2456 (2020).

[171] J. S. Yudkin, *"Prediabetes: are there problems with this label? yes, the label creates further problems!"*, Diabetes Care **39**, 1468–1471 (2016).

[172] U. P. D. S. (Group et al., *"Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (ukpds 33)"*, The lancet **352**, 837–853 (1998).

[173] J. Mathew and M. Varacallo, *"Physiology, blood plasma"*, (2018).

[174] N. L. Anderson and N. G. Anderson, *"The human plasma proteome: history, character, and diagnostic prospects"*, Molecular & cellular proteomics **1**, 845–867 (2002).

[175] F. Bonnier, M. J. Baker, and H. J. Byrne, *"Vibrational spectroscopic analysis of body fluids: avoiding molecular contamination using centrifugal filtration"*, Analytical Methods **6**, 5155–5160 (2014).

[176] D. A. Colantonio, C. Dunkinson, D. E. Bovenkamp, and J. E. Van Eyk, *"Effective removal of albumin from serum"*, Proteomics **5**, 3831–3835 (2005).

[177] S. E. Warder, L. A. Tucker, T. J. Strelitzer, E. M. McKeegan, J. L. Meuth, P. M. Jung, A. Saraf, B. Singh, J. Lai-Zhang, G. Gagne, et al., *"Reducing agent-mediated precipitation of high-abundance plasma proteins"*, Analytical biochemistry **387**, 184–193 (2009).

[178] D. J. Janecki, S. C. Pomerantz, E. J. Beil, and J. F. Nemeth, *"A fully integrated multi-column system for abundant protein depletion from serum/plasma"*, Journal of Chromatography B **902**, 35–41 (2012).

[179] C. E. McHugh, T. L. Flott, C. R. Schooff, Z. Smiley, M. A. Puskarich, D. D. Myers, J. G. Younger, A. E. Jones, and K. A. Stringer, *"Rapid, reproducible, quantifiable nmr metabolomics: methanol and methanol: chloroform precipitation for removal of macromolecules in serum and whole blood"*, Metabolites **8**, 93 (2018).

[180] E. J. Cohn, L. E. Strong, W. Hughes, D. Mulford, J. Ashworth, M. e. Melin, and H. Taylor, *"Preparation and properties of serum and plasma proteins. iv. a system for the separation into fractions of the protein and lipoprotein components of biological tissues and fluids1a, b, c, d"*, Journal of the American Chemical Society **68**, 459–475 (1946).

[181] K. M. Hosseini and M. Ghasemzadeh, *"Implementation of plasma fractionation in biological medicines production"*, Iranian journal of biotechnology **14**, 213 (2016).

[182] A. Farrugia and J. Cassar, *"Plasma-derived medicines: access and usage issues"*, Blood Transfusion **10**, 273 (2012).

[183] A. Farrugia and P. Robert, *"Plasma protein therapies: current and future perspectives"*, Best Practice & Research Clinical Haematology **19**, 243–258 (2006).

[184] J. Xiang, S. Zhang, G. Zhang, X. Li, C. Zhang, J. Luo, R. Yu, and Z. Su, *"Recovery of human serum albumin by dual-mode chromatography from the waste stream of cohn fraction v supernatant"*, Journal of Chromatography A **1630**, 461451 (2020).

[185] B. R. Don and G. Kaysen, "Poor nutritional status and inflammation: serum albumin: relationship to inflammation and nutrition", in Seminars in dialysis, Vol. 17, 6 (Wiley Online Library, 2004), pp. 432–437.

[186] S. Leto, M. J. Yiengst, C. H. Barrows Jr, et al., *"The effect of age and protein deprivation on the sulfhydryl content of serum albumin."*, Journal of gerontology **25**, 4–8 (1970).

[187] D. Perez-Guaita, S. Garrigues, et al., *"Infrared-based quantification of clinical parameters"*, TrAC Trends in Analytical Chemistry **62**, 93–105 (2014).

[188] S. Patel and S. Ahmed, *"Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery"*, Journal of pharmaceutical and biomedical analysis **107**, 63–74 (2015).

[189] A. Shevchenko and K. Simons, *"Lipidomics: coming to grips with lipid diversity"*, Nature reviews Molecular cell biology **11**, 593–598 (2010).

[190] R. Bandu, H. J. Mok, and K. P. Kim, *"Phospholipids as cancer biomarkers: mass spectrometry-based analysis"*, Mass spectrometry reviews **37**, 107–138 (2018).

[191] E. Adua, A. Russell, P. Roberts, Y. Wang, M. Song, and W. Wang, *"Innovation analysis on postgenomic biomarkers: glycomics for chronic diseases"*, Omics: a journal of integrative biology **21**, 183–196 (2017).

[192] R. Aebersold and M. Mann, *"Mass-spectrometric exploration of proteome structure and function"*, Nature **537**, 347–355 (2016).

[193] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, and M. Mann, *"Plasma proteome profiling to assess human health and disease"*, Cell systems **2**, 185–195 (2016).

[194] G. Banfi, G. L. Salvagno, and G. Lippi, *"The role of ethylenediamine tetraacetic acid (edta) as in vitro anticoagulant for diagnostic purposes"*, Clinical Chemistry and Laboratory Medicine (CCLM) **45**, 565–576 (2007).

[195] M. Wen, Y. Jin, T. Manabe, S. Chen, and W. Tan, *"A comparative analysis of human plasma and serum proteins by combining native page, whole-gel slicing and quantitative lc-ms/ms: utilizing native ms-electropherograms in proteomic analysis for discovering structure and interaction-correlated differences"*, Electrophoresis **38**, 3111–3123 (2017).

[196] A. McConnell, R. Nuttall, and D. Stalker, *"Spectroscopic studies of the metal complexes of ethylenediaminetetra-acetic acid in aqueous solution"*, Talanta **25**, 425–434 (1978).

[197] Z. Anastasiadi, G. D. Lianos, E. Ignatiadou, H. V. Harissis, and M. Mitsis, *"Breast cancer in young women: an overview"*, Updates in surgery **69**, 313–317 (2017).

[198] G. Wang, D. Zhao, D. J. Spring, and R. A. DePinho, *"Genetics and biology of prostate cancer"*, Genes & development **32**, 1105–1140 (2018).

[199] *"Lung cancer. how diet, nutrition and physical activity affect lung cancer risk"*, https://www.wcrf.org/dietandcancer/lung-cancer https://www.wcrf.org/dietandcancer/lung-cancer.

[200] W.-M. Gao, R. Kuick, R. P. Orchekowski, D. E. Misek, J. Qiu, A. K. Greenberg, W. N. Rom, D. E. Brenner, G. S. Omenn, B. B. Haab, et al., *"Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis"*, BMC cancer **5**, 110 (2005).

[201] P. Dowling, C. Clarke, K. Hennessy, B. Torralbo-Lopez, J. Ballot, J. Crown, I. Kiernan, K. J. O'Byrne, M. J. Kennedy, V. Lynch, et al., *"Analysis of acute-phase proteins, ahsg, c3, cli, hp and saa, reveals distinctive expression patterns associated with breast, colorectal and lung cancer"*, International journal of cancer **131**, 911–923 (2012).

[202] J. Lu, Y. Wang, M. Yan, P. Feng, L. Yuan, Y. Cai, X. Xia, M. Liu, J. Luo, and L. Li, *"High serum haptoglobin level is associated with tumor progression and predicts poor prognosis in non-small cell lung cancer"*, Oncotarget **7**, 41758 (2016).

[203] E. F. Patz Jr, M. J. Campa, E. B. Gottlin, I. Kusmartseva, X. R. Guan, and J. E. Herndon, *"Panel of serum biomarkers for the diagnosis of lung cancer"*, Journal of Clinical Oncology **25**, 5578–5583 (2007).

[204] M. Tolia, N. Tsoukalas, G. Kyrgias, E. Mosa, A. Maras, I. Kokakis, Z. Liakouli, J. R. Kouvaris, K. Liaskonis, N. Charalampakis, et al., *"Prognostic significance of serum inflammatory response markers in newly diagnosed non-small cell lung cancer before chemoirradiation"*, BioMed research international **2015** (2015).

[205] J.-E. S. Hansen, J. Iversen, A. Lihme, and T. C. Bøg-Hansen, *"Acute phase reaction, heterogeneity, and microheterogeneity of serum proteins as nonspecific tumor markers in lung cancer"*, Cancer **60**, 1630–1635 (1987).

[206] H. J. Johansson, F. Socciarelli, N. M. Vacanti, M. H. Haugen, Y. Zhu, I. Siavelis, A. Fernandez-Woodbridge, M. R. Aure, B. Sennblad, M. Vesterlund, et al., *"Breast cancer quantitative proteome and proteogenomic landscape"*, Nature communications **10**, 1–14 (2019).

[207] C. Mueller, A. Haymond, J. B. Davis, A. Williams, and V. Espina, *"Protein biomarkers for subtyping breast cancer and implications for future research"*, Expert review of proteomics **15**, 131–152 (2018).

[208] C. P. Tanase, E. Codrici, I. D. Popescu, S. Mihai, A.-M. Enciu, L. G. Necula, A. Preda, G. Ismail, and R. Albulescu, *"Prostate cancer proteomics: current trends and future perspectives for biomarker discovery"*, Oncotarget **8**, 18497 (2017).

[209] P. Intasqui, R. P. Bertolla, and M. V. Sadi, *"Prostate cancer proteomics: clinically useful protein biomarkers and future perspectives"*, Expert review of proteomics **15**, 65–79 (2018).

[210] X. Wang, X. Shen, D. Sheng, X. Chen, and X. Liu, *"Ftir spectroscopic comparison of serum from lung cancer patients and healthy persons"*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **122**, 193–197 (2014).

[211] V. E. Sitnikova, M. A. Kotkova, T. N. Nosenko, T. N. Kotkova, D. M. Martynova, and M. V. Uspenskaya, *"Breast cancer detection by atr-ftir spectroscopy of blood serum and multivariate data-analysis"*, Talanta **214**, 120857 (2020).

[212] D. L. Freitas, I. M. Câmara, P. P. Silva, N. R. Wanderley, M. B. Alves, C. L. Morais, F. L. Martin, T. B. Lajus, and K. M. Lima, *"Spectrochemical analysis of liquid biopsy harnessed to multivariate analysis towards breast cancer screening"*, Scientific Reports **10**, 1–8 (2020).

[213] D. K. Medipally, A. Maguire, J. Bryant, J. Armstrong, M. Dunne, M. Finn, F. M. Lyng, and A. D. Meade, *"Development of a high throughput (ht) raman spectroscopy method for rapid screening of liquid blood plasma from prostate cancer patients"*, Analyst **142**, 1216–1226 (2017).

[214] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, *"Progress and prospects of early detection in lung cancer"*, Open biology **7**, 170070 (2017).

[215] D. C. McMillan, W. S. Watson, P. O'Gorman, T. Preston, H. R. Scott, and C. S. McArdle, *"Albumin concentrations are primarily determined by the body cell mass and the systemic inflammatory response in cancer patients with weight loss"*, Nutrition and cancer **39**, 210–213 (2001).

[216] D. Gupta and C. G. Lis, *"Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature"*, Nutrition journal **9**, 69 (2010).

[217] S. W. Merriel, R. Carroll, F. Hamilton, and W. Hamilton, *"Association between unexplained hypoalbuminaemia and new cancer diagnoses in uk primary care patients"*, Family practice **33**, 449–452 (2016).

[218] A. Tastanova, M. Folcher, M. Müller, G. Camenisch, A. Ponti, T. Horn, M. S. Tikhomirova, and M. Fussenegger, *"Synthetic biology-based cellular biomedical tattoo for detection of hypercalcemia associated with cancer"*, Science translational medicine **10**, eaap8562 (2018).

[219] J. Raynes, S. Eagling, and K. McAdam, *"Acute-phase protein synthesis in human hepatoma cells: differential regulation of serum amyloid a (saa) and haptoglobin by interleukin-1 and interleukin-6"*, Clinical & Experimental Immunology **83**, 488–491 (1991).

[220] T. Hochepied, F. G. Berger, H. Baumann, and C. Libert, *"α 1-acid glycoprotein: an acute phase protein with inflammatory and immunomodulating properties"*, Cytokine & growth factor reviews **14**, 25–34 (2003).

[221] V. Arroyo, R. Garcıa-Martinez, and X. Salvatella, *"Human serum albumin, systemic inflammation, and cirrhosis"*, Journal of hepatology **61**, 396–407 (2014).

[222] J. F. Salgado, *"Transforming the area under the normal curve (auc) into cohen'sd, pearson's rpb, odds-ratio, and natural log odds-ratio: two conversion tables"*, European Journal of Psychology Applied to Legal Context **10**, 35–47 (2018).