POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Raman imaging of patient-derived tumour slice cultures: towards personalized therapy monitoring

Tesi di Laurea Magistrale in
Engineering Physics - Ingegneria Fisica

Author: **Chiara Agrimi**

Student ID: 10683815
Advisor: Prof. Dario Polli
Co-advisors: Renzo Vanna, Victor Alcolea-Rodriguez
Academic Year: 2024-25

Abstract

In modern oncology, the selection of an effective therapy for a patient still presents significant challenges due to the complexity and variability of the disease. In spite of the efforts, innovation and continuous improvements, the success rates of therapeutic choices remain suboptimal in many cases; the ability to monitor treatment response early in time and predict accurately patient-specific outcomes represents therefore an urgent and unmet need in clinical practice.

Conventional approaches for therapy monitoring, such as imaging techniques, are often limited by late feedback, while immunohistochemistry for evaluation of tumor microenvironment lacks of quantitative references and threshold for resolutive assessments.

In this framework, Raman spectroscopy emerges as a promising technique, offering label-free, non-destructive, and chemically specific insights into biological samples. In particular, this thesis investigates the use of spontaneous Raman spectroscopy applied to patient-derived head and neck squamous cell carcinoma (HNSCC) slice cultures.

A home-built Raman microspectroscopy system was employed to acquire hyperspectral maps of HNSCC samples subjected to different treatment conditions on different timepoints. A dedicated preprocessing workflow was designed, together with custom Python tools for background subtraction, artifact identification and pixel-wise annotation of maps. Statistical analyses were subsequently performed using both univariate and multivariate methods.

The results demonstrate that Raman spectroscopy is able to capture biochemical differences stemming from different tissue composition or associated with therapeutic interventions. By means of principal component analysis, providing programmed death ligand 1 (PDL1) protein was found to induce detectable modifications in DNA and protein-related Raman bands, while clustering approaches highlighted the possibility of discriminating tumor, tumor stroma, skeletal muscle and other healthy tissue structures across patients. These findings confirm the validity of Raman spectroscopy as a tool for the characterization ex-vivo patient-derived slice cultures, with the potential of revealing insights also on changes produced by medical treatment occurring in time.

Keywords: Spontaneous Raman imaging, patient-derived tumour slice cultures, tissue characterization, therapy monitoring.

Abstract in lingua italiana

Nell'oncologia moderna, la selezione di una terapia efficace per i pazienti presenta sfide rilevanti a causa della complessità e della variabilità della malattia. Nonostante le innovazioni e i continui progressi, i tassi di successo rimangono in molti casi non ottimali; la capacità di monitorare tempestivamente la risposta al trattamento e prevedere con precisione gli esiti specifici per un paziente rappresenta quindi una necessità urgente e ancora insoddisfatta in ambito clinico.
Gli approcci convenzionali per il monitoraggio di terapia, come le tecniche di imaging, sono spesso limitati da feedback tardivi, mentre l'immunoistochimica per la valutazione del microambiente tumorale manca di riferimenti quantitativi e soglie per valutazioni risolutive. In questo contesto, la spettroscopia Raman emerge come una tecnica promettente, caratterizzata da specificità chimica, non distruttività e assenza di label per marcare i campioni biologici. In particolare, questa tesi indaga l'uso della spettroscopia Raman spontanea applicata a colture di sezioni di carcinoma squamoso della testa e del collo (HNSCC) derivate da paziente.
Un sistema di microscopia Raman sviluppato ad-hoc è stato utilizzato per acquisire mappe iperspettrali di campioni HNSCC sottoposti a diverse tipologie di trattamento e a diversi istanti temporali. È stata progettata una pipeline di preprocessing, insieme a codici Python personalizzati per la sottrazione del background, l'identificazione di artefatti e l'annotazione pixel-per-pixel delle mappe. Successivamente sono state eseguite analisi statistiche con approcci univariati e multivariati.
I risultati dimostrano che la spettroscopia Raman è in grado di cogliere differenze biochimiche derivanti dalla diversa composizione tissutale o associate a interventi terapeutici. Tramite l'analisi delle componenti principali , è stato osservato che il trattamento con proteina PDL1 induce cambiamenti nelle bande Raman relative a DNA e proteine, mentre approcci di clustering hanno evidenziato la possibilità di discriminare tumore, stroma tumorale, muscolo scheletrico e altre strutture tissutali sane tra i diversi pazienti. Questi risultati confermano la validità della spettroscopia Raman come strumento per la caratterizzazione di colture di sezioni derivate da pazienti, con il potenziale di rivelare anche cambiamenti nel tempo indotti da un eventuale trattamento.

Parole chiave: Imaging Raman spontaneo, Colture di sezioni tumorali da paziente, caratterizzazione di tessuti, monitoraggio di terapia.

Contents

Abstract **i**

Abstract in lingua italiana **iii**

Contents **v**

1 Introduction **1**

1.1 Raman scattering 1

1.1.1 Theoretical treatment - Classical model 2

1.1.2 Theoretical treatment - Semi-classical model 7

1.1.3 Spontaneous Raman for biomedical applications 9

1.2 Assessment of patient-specific cancer therapy responses: an urgent clinical need 12

1.2.1 Introduction to cancer and cancer therapies 12

1.2.2 Current approaches for therapy monitoring and prediction 16

1.2.3 Case study: HNSCC 19

1.3 Extracting information from Raman spectra: preprocessing and data analysis tools 22

1.3.1 Preprocessing 22

1.3.2 Data analysis 25

2 Aim of the study **29**

3 Materials and methods **31**

3.1 Experimental setup 31

3.2 Samples description 34

3.3 Data collection 38

3.3.1 Measurements and acquisition parameters 38

3.3.2 Wavenumber and intensity calibration 39

3.3.3 Data preprocessing 41

4 Experimental results and discussion **47**

4.1 Design and optimization of the preprocessing pipeline 47

4.1.1 Tested pipelines and comparison 47

4.1.2 Custom made preprocessing tools 57

4.2 Raman maps analysis 67

4.2.1 Analysis of average foreground spectra 69

4.2.2 PCA analysis 78

4.2.3 K-means clustering for tissue type identification 87

5 Conclusions and future developments **103**

Bibliography **105**

A Appendix - Map nomenclature and dataset for the analysis **111**

List of Figures **113**

List of Tables **115**

Acknowledgements **117**

1 | Introduction

The work presented in this thesis is focused on spontaneous Raman microspectroscopy applied for imaging patient-derived tumor slice cultures, specifically for the cancer category of head and neck squamous cell carcinomas (HNSCC). The main goal of the activities will be the biochemical investigation and characterization of the measured samples, in the framework of a larger research project aiming at using Raman and multimodal imaging techniques to identify biomarkers of tumor resistance, introducing a label-free, nondestructive and rapid methodology to tackle the issue related to cancer therapy monitoring and outcome prediction.
A brief introduction on the fundamental principles of the Raman scattering phenomenon will be given, with remarks to the characteristics that make the technique particularly attractive in the field of biomedical studies. A general overview regarding cancer, current therapeutic approaches and therapy assessment methodologies will follow, with additional details on HNSCC tumors specifically investigated in this work. The scope of these brief descriptions is contextualizing the activities of this work, introducing at the same time the critical clinical need that the project is aiming to address.
The actual challenges and tools for spectral data preprocessing and analyses will conclude the introduction to this work, before proceeding with the illustration of the methodology adopted and the discussion of experimental results.

1.1. Raman scattering

When light interacts with matter, two different phenomena may occur: absorption and scattering. In the first case, the energy carried by radiation is stored in the material, and eventually dissipated, while in the latter light is transmitted either along its initial propagation direction or on a different, deflected one. Most of the radiation deriving from scattering is characterized by the same frequency of the incident one; in this case, the phenomenon is referred to as Rayleigh scattering (or elastic scattering, since the energy of the light is preserved). However, it is possible to also identify a portion of light that is instead frequency-shifted with respect to the impinging one, and in that case this inelastic

process is known as Raman effect or Raman scattering, owing the name to the Indian physicyst Chandrasekhara Venkata Raman, who was the first to report this phenomenon and was awarded the Nobel Prize in Physics in 1930 for his discovery. The scattered light can either have an higher or a lower frequency with respect to the incident one, and the two cases are respectively known with the name of "Stokes scattering" and "anti-Stokes scattering".

To better understand the working principle and the applications of Raman spectroscopy, it is essential to explore the theoretical aspects of the Raman scattering process with more detail. In the following paragraph the concept of molecular vibrational levels and the Raman scattering process will be presented, both according to two different yet equivalent frameworks: classical physics and quantum theory.

1.1.1. Theoretical treatment - Classical model

In the classical treatment, the behaviour of the chemical link binding together two atoms in a molecule is often resembled to the one of a spring, characterized by a rest length corresponding to the equilibrium distance between the two atoms, and a stiffness coefficient K, which depends on the bond strength (energy), the equilibrium bond length, the bond order (single, double, triple), and the size and electronic nature of the atoms involved. In this framework, supposing that the two atoms are displaced by a distance x_1 and x_2 from their equilibrium position, the force F applied on the two atoms due to the chemical bond, according to Hooke's law, will be equal to:

$$F = -K(x_1 + x_2) \tag{1.1}$$

Being m_1, m_2 and r_1, r_2 the masses of the two atoms and their positions with respect to the center of masses respectively, such that $r_1 + r_2 = r_0$ rest length of the chemical bond, one can get that:

$$\begin{cases} m_1 r_1 = m_2 r_2 \\ m_1(r_1 + x_1) = m_2(r_2 + x_2) \end{cases} \quad \Rightarrow \quad x_1 = \frac{m_2}{m_1} x_2 \tag{1.2}$$

$$F = -K\left(\frac{m_1 + m_2}{m_1}\right) x_2 \tag{1.3}$$

On the other side, the second law of dynamics imposes that

$$F = ma \tag{1.4}$$

Since the only force acting on the two atoms is the elastic one of the chemical bond, one can equate the right-hand sides of eq 1.3 and 1.4 for both atoms 1 and 2. By proper manipulation of the resulting equations and introducing the reduced mass M and the displacement q as indicated below, the equation can be rewritten in the following form:

$$M\frac{d^2q}{dt^2} + Kq = 0 \tag{1.5}$$

with

$$M = \frac{m_1 m_2}{m_1 + m_2} \tag{1.6}$$

$$q = x_1 + x_2 \tag{1.7}$$

which has the following solutions

$$q = q_0 sin(2\pi\nu_0 t + \phi) \tag{1.8}$$

with

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{K}{M}} \tag{1.9}$$

This means that the molecule can oscillate with an harmonic oscillation of amplitude q_0 and frequency v_0, with the latter entailing information on the molecule itself. In particular, the frequency v_0 depends only on the reduced mass M which provides information on the kind of atoms forming the molecule (each element is associated with its mass) and the stiffness of the bond, which is directly linked to its strength. Strong bonds such as triple bonds usually feature high values of K, resulting in higher oscillation frequencies of the related mode. The whole treatment can then be generalized considering a 3D model of the system and polyatomic molecules composed by a generic N number of atoms; in that case, considering the 3 degrees of freedom of each atom and the null solutions due to the symmetry of the system, a total of 3N-6 normal modes of oscillations at 3N-6 frequencies can be found (in case of non-degenerate solutions).

Considering now the interaction of an electromagnetic field with a molecule or an atom, the oscillating electric field E impinging on a sample causes the oscillation of electric charges

inside the sample, thus producing an induced dipole moment that, in first approximation, is characterized by a linear dependence on the electric field:

$$\vec{\mu} = \alpha \vec{E} \tag{1.10}$$

where α is a 3x3 tensor named polarizability. The polarizability of a molecule is influenced by its vibrational motion; supposing small-amplitude oscillations, we can express this dependence with a Taylor series expansion around the equilibrium value of each normal coordinate. To further simplify the treatment, we will only consider the first term (linear) of the expansion, which is a reasonable approximation as long as we are considering small amplitude vibrations near the equilibrium. We finally obtain, for each normal vibrational coordinate:

$$\alpha_k = \alpha_0 + \left(\frac{\delta \alpha}{\delta q_k}\right)_0 q_k \tag{1.11}$$

Considering a normal mode k, introducing in 1.10 the expression of the oscillating mode 1.8 and considering a sinusoidally oscillating electric field at an oscillating frequency ν, one can obtain the overall expression of the induced dipole moment:

$$\mu = \alpha_0 E_0 \sin(2\pi c\nu t) + \left(\frac{\partial \alpha}{\partial q_k}\right)_0 q_{k0} E_0 \cos(2\pi c\nu t) \cos(2\pi c\nu_k t) \tag{1.12}$$

or

$$\mu = \alpha_0 E_0 \cos(2\pi c\nu t) + \frac{1}{2}\left(\frac{\partial \alpha}{\partial q_k}\right)_0 q_{k0} E_0 \left[\cos 2\pi c(\nu - \nu_k)t + \cos 2\pi c(\nu + \nu_k)t\right] \tag{1.13}$$

The second expression is equivalent to the first one and is obtained by applying trigonometric transformation. The equation 1.13 represents a particularly useful formulation since it allows the identification of different scattering phenomena occurring within the material. The induced dipole moment is composed in fact by:

1. A term oscillating with a frequency equal to the one of the excitation light. It corresponds to the Rayleigh scattering radiation, for which the impinging light undergoes an elastic scattering process and is deflected without any change in its energy;

2. A term oscillating at a red-shifted frequency with respect to the excitation one, with a shift equal to the frequency of oscillation of one normal mode $(\nu - \nu_k)$. This term corresponds to the Stokes Raman signal, and can be explained as a portion of the impinging light energy to be absorbed by the molecule, with the excitation

(oscillation) of the k-th normal mode;

3. A term oscillating at a blue-shifted frequency with respect to the excitation one, again with a shift equal to the frequency of oscillation of one normal mode ($\nu + \nu_k$). This component corresponds to the anti-Stokes Raman signal.

It is worth noticing that the Stokes and anti-Stokes signals are also characterized by a phase term ϕ which varies from one scattering event to the other; this makes the Raman process intrinsically incoherent, resulting in a lower intensity of the output Raman signal due to the incoherent superposition of the contributions from different molecules. Furthermore, the anti-Stokes light is characterized by an increase in frequency, meaning that the energy is transferred from the sample to the scattered light. This phenomenon can occur only in the case in which the scattering molecule is already in an excited state, with the k-th mode oscillating, which then transfers its energy to the scattered light. However, this condition is relatively uncommon; in fact, as Boltzmann distribution suggests, the probability of finding the molecule at an higher energy with respect to the rest energy decreases exponentially with the energy of the excited vibrational level:

$$p(E) = e^{-\frac{E}{k_B T}} \tag{1.14}$$

As a result, anti-Stokes peaks typically exhibit lower intensity compared to Stokes peaks, and approach the height of the latter only in case of very low energy vibration modes, or when the sample has a high thermal energy.

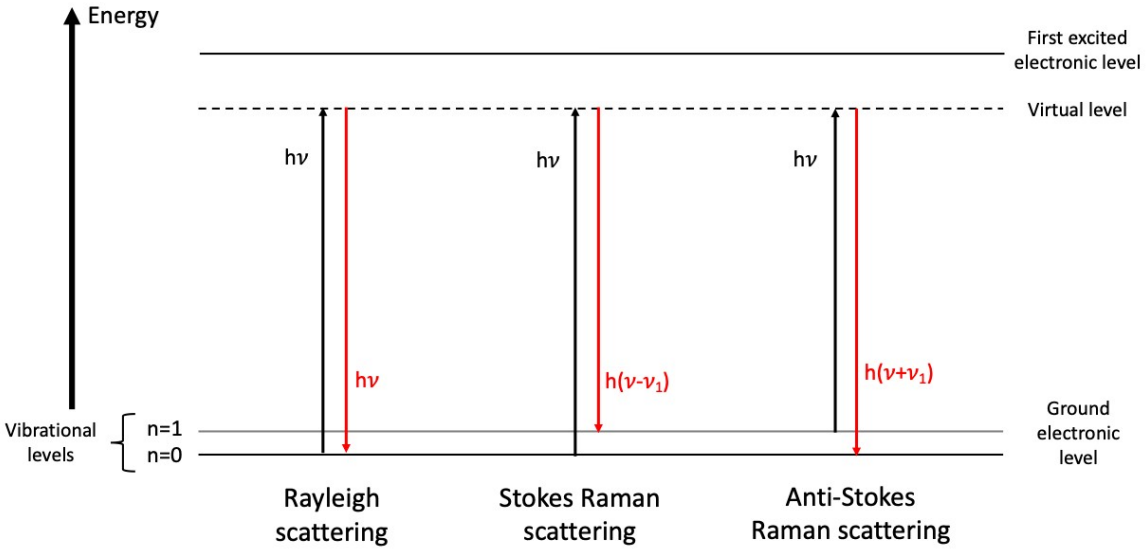


Figure 1.1: **Schematic representation of the possible scattering interactions**.

Since the system behaves like an oscillating dipole, it will emit an electromagnetic radiation, which is in fact the scattered light. Considering a certain direction of propagation forming an angle θ with respect to the oscillating dipole, the electric field amplitude E_0 at a given distance r from the dipole will be:

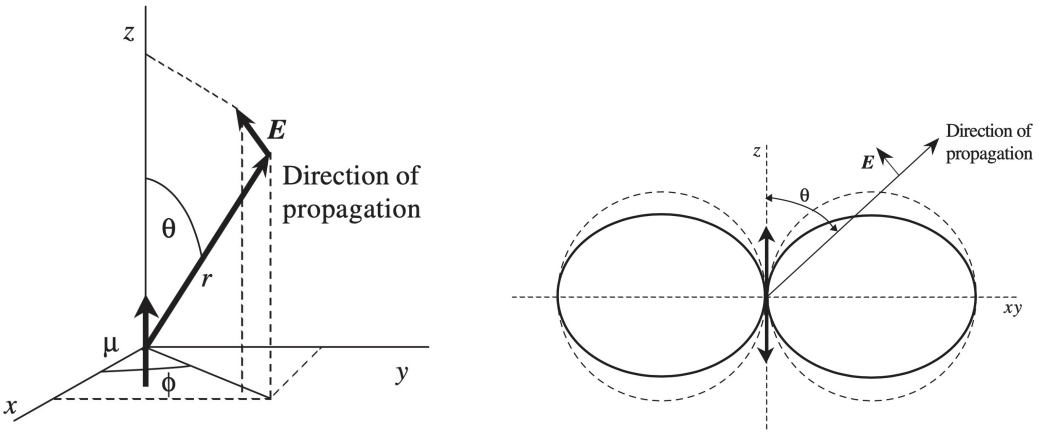
$$E_0 = \frac{\pi \mu_0 \nu^2 \sin\theta}{\varepsilon_0 r} \tag{1.15}$$

where ν is the wavenumber of the radiation, m_0 is the magnitude of the oscillating dipole, ε_0 is the permittivity of the medium. The corresponding time average energy density is:

$$\rho = \frac{\pi^2 \mu_0^2 \nu^4 \sin^2\theta}{2\,\varepsilon_0 r^2}. \tag{1.16}$$

In figure 1.2 the angular distribution of emitted radiant intensity is shown; from both the plot and the formula, one can observe that the radiant intensity has its maximum at the equatorial plane $\theta = \frac{\pi}{2}$ while decreases to 0 when θ approaches 0.

Integrating on the whole solid angle, one can compute the overall power Φ, useful to introduce a further parameter to describe the efficiency of the process: the scattering cross section, defined as the ratio between Φ and the irradiance I of the impinging radiation:



(a) (b)

Figure 1.2: **Angular dependence of the emitted radiant intensity of an oscillating dipole [1].**

$$\sigma = \frac{\Phi}{I} \tag{1.17}$$

Finally, from the previous expression 1.13, it is possible to see that not all normal modes of vibration necessarily produce a Raman scattering signal; indeed, if the polarizability α is independent from the vibrational coordinate q_k, the derivative in 1.11 cancels out and the only term contributing to μ is the one with frequency equal to the incident one, i.e. corresponding to Rayleigh scattering. The condition:

$$\left(\frac{\partial \alpha}{\partial q_k}\right)_0 = 0 \tag{1.18}$$

is often referred to as the transition selection rule for the Raman vibration corresponding to the oscillation of q_k.

1.1.2. Theoretical treatment - Semi-classical model

The concepts of vibrational levels and Raman scattering can also be described using quantum mechanics; in this specific framework, it is useful to adopt a semi-classical approach, for which radiation is still treated according to classical physics, while molecules are described by means of quantum mechanics, and the light-matter interaction is considered as a small perturbation of the equilibrium condition of the system, thus allowing to study the phenomenon of Raman scattering with a perturbative approach. According to quantum mechanics, the Schrodinger equation in normal coordinates for the system described in

the previous paragraph will be the following:

$$\frac{\hbar^2}{2\mu}\frac{d^2\Psi}{dq^2} + \left(-\frac{1}{2}Kq^2\right)\Psi = E\Psi \tag{1.19}$$

where $V = -\frac{1}{2}Kq$ is the potential which the system is subject to, that in our treatment includes only the elastic energy existing due to the chemical bond. The eigenvalues of equation 1.19 are:

$$E_n = h\nu\left(n + \frac{1}{2}\right) \tag{1.20}$$

Where

$$\nu = \frac{1}{2\pi}\sqrt{\frac{K}{\mu}} \tag{1.21}$$

yielding the same result as before, but introduces an additional concept that cannot be captured within a classical framework, that is the quantization of energy. From the quantum treatment in fact we see that the energy of vibrational levels is not a continuous quantity but can only assume discrete values, that are multiples of the quantity $\frac{1}{2}h\nu$. As a consequence, in the light-matter interaction, only discrete packets of energy can be exchanged between the radiation and the molecule, and the minimum energy of the system, which is obtained by imposing n=0, is not null but equal to $\frac{1}{2}h\nu$.

Supposing now to have an impinging electric field on the molecule with moderate intensity, the change produced on the system can be described according to a perturbative approach (perturbation theory). The perturbation Hamiltonian has the same expression as its classical counterpart:

$$H = -E \cdot \hat{M} \tag{1.22}$$

where E is the incident electric field and $\hat{M}$ is no more the electric dipole moment of the molecule but rather the corresponding operator. According to the perturbation theory, the probability for having a transition from an initial stationary state Ψ_i to a final stationary state Ψ_f is proportional to the square modulus of the transition dipole moment element μ_{fi}, calculated as:

$$\mu_{fi} \;=\; \langle\Psi_f|\hat{\mu}|\Psi_i\rangle \;=\; \int \Psi_f^*\,\hat{\mu}\,\Psi_i\,d\tau \tag{1.23}$$

$$P_{trans} \propto |\mu_{fi}|^2 \tag{1.24}$$

By performing the computation, one finally obtains that the matrix element μ_{fi} differs from zero only for transitions for which the final state and the initial one differ only by one vibrational quantum number: for Stokes signals $n_f = n_i + 1$, while for anti-Stokes signals $n_f = n_i - 1$. Considering single photon interactions, μ_{fi} will be different from zero only for one specific vibrational mode k, meaning that only one vibrational mode can be excited by one photon. However, this does not mean that, in principle, a single molecule cannot reach 'higher' vibrational levels, but rather that these higher levels cannot be achieved through single photon interaction events. The selection rule $\Delta n_k = \pm 1$ for one and only one vibrational mode k is derived under the condition of double harmonic approximation; in real cases, anharmonicity lead to a relaxation of this constraint, with overtones and combination of the fundamental vibrational frequencies may appear. The requirement 1.18 still holds as necessary condition to have Raman scattering.

One final important difference compared to classical treatment is in relative intensities of Stokes and anti-Stokes peaks, whose ratio corresponds to:

$$\frac{I_{\text{Stokes}}}{I_{\text{anti-Stokes}}} = \frac{I_{\nu-\nu_k}}{I_{\nu+\nu_k}} = \frac{(\nu-\nu_k)^4}{(\nu+\nu_k)^4} \exp\left(\frac{hc\nu_k}{kT}\right) \tag{1.25}$$

Following a classical approach, this difference is not found, but the lower height of anti-Stokes peaks is explained by means of Boltzmann statistics as previously illustrated.

1.1.3. Spontaneous Raman for biomedical applications

Raman scattering has an intrinsic high chemical specificity: the vibrational frequencies of oscillation are strongly related to the atomic composition of molecules and the kind of bonds that keep them together; in this framework, one could say that the set of Raman transitions frequencies that constitute a molecule's Raman spectrum of a molecule can be considered a fingerprint of the molecule itself, identifying it uniquely. Combined with the recent broad technological advances, this property has established Raman spectroscopy as a very powerful analytical technique, which is nowadays widely applied across diverse scientific disciplines, including chemistry, physics, materials science, pharmacy, biology, biomedicine, geology, mineralogy, and environmental science [2].

In the specific framework of biomedical studies, together with the abovementioned chemical selectivity, Raman spectroscopy and microscopy offer many other benefits:

- Minimal sample preparation;
- Non-destructive measurements preserve the integrity of the sample, allowing it to be studied without alteration and enabling repeated acquisitions or complementary analyses. This addresses a major limitation of immunohistochemistry and sectioning techniques.;
- Label-free, allowing the study and false-color imaging of samples without the need of additional substances or markers, required instead in fluorescence microscopy. The use of fluorescent tags poses toxicity-related issues, and in some cases may be difficult to implement or can interfere with the biological functions of the sample;
- Rapid acquisition, especially in the context of intra-operative applications and for surgical-aid devices.

An example of Raman spectrum is provided in figure 1.3, where each peak corresponds to a particular oscillation mode. It is fundamental to note that the elements and types of bonds that characterize biomolecules are such that the associated oscillation frequencies fall within the ranges of 0-1800 cm^{-1} and 2800-3500 cm^{-1}, while no peaks are usually found in the wavenumber range of 1800-2800 cm^{-1}. It is therefore common to refer to the various wavenumber regions according to the following nomenclature:

Fingerprint region (500 - 1800 $\boldsymbol{cm^{-1}}$), where one can find peaks associated to a large variety of components, including lipids, proteins, phosphate groups of DNA and RNA, metabolites, aminoacids, Amide I and Amide III band. The fingerprint region is very sample specific and particularly suitable to uniquely identify the sample thanks to its chemical composition.

Silent region (1800 - 2800 $\boldsymbol{cm^{-1}}$), which contains peaks associated to triple bonds such as $C \equiv C$ and $C \equiv N$ and to other single or double bonds which are not present in biomolecules, thus the name of 'silent region'.

CH region (2800 - 3500 $\boldsymbol{cm^{-1}}$), that owes its name since it is specifically contributed by the stretching vibrations of CH groups, predominantly present in lipids and proteins [3].

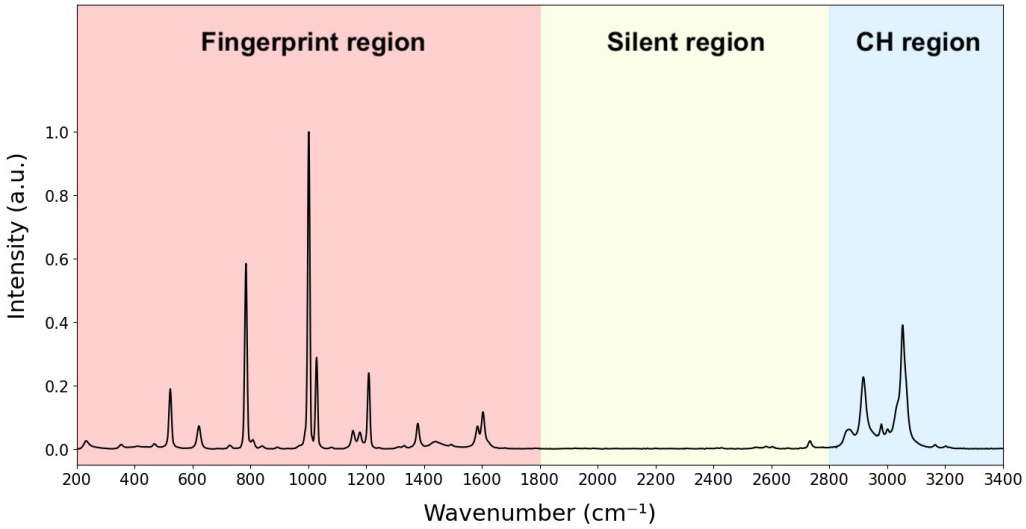


Figure 1.3: **The three spectral regions (fingerprint, silent and CH) on a sample Raman spectrum (toluene)**.

On the other side, Raman scattering suffers as previously mentioned from a low cross section, which significantly hinders the sensitivity unless this aspect is counterbalanced with long exposure time measurements; in addition, Raman scattering competes with other absorption-emission mechanisms such as autofluorescence, which produces a broadband emission of light at frequencies below the excitation one which deforms the Raman spectrum and complicates or even makes impossible the extraction of Raman features. Nevertheless, different approaches have been adopted in order to overcome these limitations, leading to the creation of a large variety of Raman spectroscopy techniques; to cite few of them, spatially offset Raman spectroscopy (SORS) is characterized by a more complex geometry but allows probing deeper tissue layers than standard Raman; surface enhanced Raman spectroscopy (SERS) requires the use of metallic probes but benefits of a significant increase in Raman signal thanks to the deriving photonic interactions; coherent Raman spectroscopy techniques, such as SRS and CARS, exploit non-linear optical phenomena to obtain an enhanced signal intensity, up to six orders of magnitude larger than the spontaneous one [4], allowing much more rapid acquisition at some expenses in terms of information content of each measurement (not a full Raman spectrum but single-wavelength signals).

Numerous examples exist in the literature of Raman spectroscopy applications in medicine, particularly for fundamental research, diagnostics, classification, and technological applications in clinics for the support of healthcare professionals. We will now focus specifically on the oncology sector, starting with a general overview before directing our attention to

an urgent and strongly felt need: a fast, label-free, and accurate tool designed to predict the efficacy of anticancer therapies and to monitor their outcomes.

1.2. Assessment of patient-specific cancer therapy responses: an urgent clinical need

Cancer is nowadays recognized as one of the most crucial public health challenges. According to GLOBOCAN statistics, cancer is responsible for almost one in six deaths (16.8%) and one in four deaths (22.8%) from noncommunicable diseases (NCDs) worldwide. For the year 2022 almost 20 million new cases of cancer were registered, and close to 10 million people lost their lives to this disease. In addition, demographics-based predictions indicate that the annual number of new cases of cancer is expected reach 35 million by 2050 [5]. The scientific community is making great collective effort to find effective cancer cures. Yet, the choice of an appropriate anti-cancer therapy is still a very critical aspect, being strongly patient dependent: many factors indeed are to be taken in account, balancing between effectiveness of the therapy and potential negative side effects caused to the patient.

The extreme complexity of the problem and the extensive number of techniques that have been developed do not allow a fully comprehensive discussion here. In the following paragraphs, a brief introduction on cancer, cancer therapies, and established monitoring approaches will be provided, with the sole purpose of outlining the context of this work and the clinical need driving it.

1.2.1. Introduction to cancer and cancer therapies

The term cancer refers to a diverse and numerous group of diseases for which some cells grow uncontrollably and eventually spread to other parts of the body. The process of cancer generation (carcinogenesis) can be linked to mutagenesis, that is the production of changes in the DNA sequence, especially in the cases of exposure to external chemical agents and radiation such as X-rays and UV light. Random alteration of the DNA can also occur by chance during healthy cells division and may lead to the development of a cancer if it is not fixed by innate DNA repair mechanisms. For the same reason, individuals inheriting genetic defects that affect DNA repair mechanisms are exposed to much higher risks. Many different types of cancer exists, each distinguished by a variety of characteristics, such as cellular origin, histological features, genetic mutations and growth patterns. These distinctions are closely linked to the anatomical site where the cancer

first develops, which significantly influences both the biological behavior of the tumor and its clinical management. Nevertheless, some common key attributes are generally found in cancer cells in contrast to healthy ones [6]:

1. an altered homeostasis that results in cells growing and dividing at a faster rate than they die;
2. the bypassing of normal limits to cell proliferation;
3. evasion of cell-death signals;
4. altered cellular metabolism;
5. manipulation of the tissue environment to support cell survival and to evade a deleterious immune response;
6. escape of cells from their home tissues and proliferation in foreign sites (metastasis);

Additionally, tumoral masses are usually characterized by an abnormal angiogenesis, with the formation of new blood vessels that allows the malignant cells to be provided with oxygen and nutrients to sustain their proliferation and tumor expansion [7].
The peculiar characteristics of each type of cancer, the variability shown from patient to patient, the complexity of the disease and the subtle mechanisms that it acts for bypassing healthy biological control processes are all factors that enormously complicate the treatment of this condition.

Nowadays, the most common and widespread choice for cancer treatment consists in surgical removal of the tumoral mass, radiotherapy, chemotherapy, immunotherapy or combinations of them.
The **excision** of the tumoral mass is one of the most straightforward approaches for localized cancer tissues; however, this option is practically not possible in case of non-solid tumors and its feasibility depends on the characteristics of the tumor itself, such as its stage of development, extension and localization: in some conditions, such as tumors affecting brain, nervous system or extensive areas comprising vital organs, the surgical removal is often to be excluded because it may be too dangerous and might cause permanent if not fatal damage on the patient. In this framework, it is important to remember that in the majority of cases the excision involves also the removal of a small healthy tissue layer around the tumoral mass, to ensure the absence of residual tumoral cells in the patient that can lead to a relapse of the tumor. When this is not possible for one of the abovementioned reason or due to unclear tumor margins, surgery may not be resolutive. Finally, surgery is excluded when multiple metastases have been created or when the patient is considered to have an elevated surgical risk, with significantly higher mortality

and complication rates linked, for example, to the presence of comorbirdities.

Chemotherapy: Chemotherapic approaches consist in the introduction of toxic substances with the aim of targeting and killing cancer cells. These drugs can can suppress tumor growth through diverse mechanisms, for example working on crucial cellular enzymes, altering cell metabolism, or interfering with some critical cellular processes, such as programmed cell death (apoptosis), drug resistance, DNA damage, DNA replication, or immune reactions [8]. The idea standing for this approach is based on what has already been mentioned for radiotherapic approaches: the stress condition under which tumoral cells live lowers their capability to deal with other mechanisms altering the cell equilibrium; as a consequence, the therapy is expected to be more effective in damaging and eliminating cancer cells. An example of chemotherapeutic agents are Alkylating agents and platinum-based agents, which, even if with two different mechanisms, keep the cell from reproducing by damaging its DNA. Other cathegories include antimetabolites, mitotic inhibitors, topoisomerase inhibitors, and anti-tumors antibiotics (targeting DNA of cancer cells).

Immunotherapy aims at enhancing the ability of the immune system to act against tumor cells. Some relevant examples of this kind of therapy include:

1. Immune checkpoint inhibitors: T-cells of the immune system are provided with membrane proteins, called immune checkpoint proteins, that can bind to complementary proteins expressed on the surface of healthy cells. When the interaction between partner immune checkpoint proteins occurs, the action of the T-cells is regulated towards a reduction of the immune system response, preventing it from being excessive and potentially harmful. In some cases, cancer cells are able to express those complementary proteins too, with the result that, even if cytotoxic T-cells recognize tumor antigens, they are prevented from killing those cells. An example of this is given by the programmed cell death ligand 1 (PD-L1), which, by binding to PD1 (programmed cell death 1) located on T cells, induces a series of mechanisms and processes that promote T-cell apoptosis, with the overall effect of reducing the immune response. When T-cells encounter cancer cells exhibiting PD-L1 protein on their surface, PD-L1 binds to PD-1 and T-cells respond in a similar way as if they had recognized the presence of an healthy cell. In order to fully exploit the immune system capability to combat cancer cells, specific anticancer drugs called immune checkpoint inhibitors (ICI) can take the place of the partner protein, preventing the binding of PD-1-PD-L1 between T-cells and cancer cell and disrupting in this way the mechanism allowing tumor cells to evade the immune system. Another example is the targeting of Cytotoxic-T-lymphocyte-antigen-4 (CTLA-4), which plays a sim-

ilar role in the suppression of the immune system response and has been reported to be overexpressed in malignant cells of different cancer types [9].

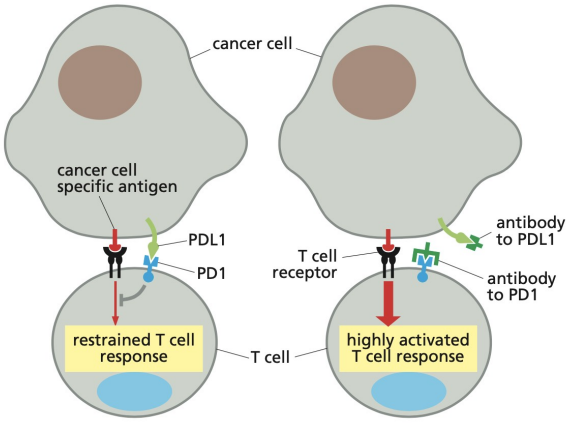


Figure 1.4: **Schematic example of how cancer cells protect themselves from immune response**: the cancer cell expresses PD-L1, which binds the PD-1 receptor on a T cell and interferes with T cell activation. This makes the tumor susceptible to antibodies that unleash T cell attack [6].

2. Monoclonal antibodies: monoclonal antibodies are laboratory-engineered proteins that replicate the function of naturally occurring antibodies. In particular, they can bind selectively to defined antigens, including those expressed by cancer cells, and can be used as 'markers' that facilitate the identification and detection from the B-cells and T-cells of the immune system. For examples, monoclonal antibodies like Trastuzumab are easily recognized by immune cells expressing Fc receptors (macrophages, natural killer cells, dendritic cells); Trastuzumab, on the other side, can bind to HER2, a protein expressed on the surface of some types of cancers like gastric and breast cancer. The result is that immune cells recognize the monoclonal antibody bound on the cancer cells' surface and are activated to eliminate it.

3. T-cell Transfer Therapy: it consists in the extraction of T-cells that are found in the patient's tumor, followed by a selection depending on their ability to recognize the tumor. Cells that are more able to identify the tumor are cultured and grown in large numbers, to be finally injected back inside the patient with the aim of boosting the immune response. When the extracted T-cells are the ones located inside or near the tumor (tumor-infiltrating lymphocytes, or TIL) the therapy takes the name of TIL therapy. CAR T-cell therapy is a variant of this, for which TIL are modified before re-injection so to be able to produce protein known as chimeric antigen receptor (CAR).

Radiotherapy: Radiotherapy employs ionizing radiation that causes damages to the DNA of cells and interfere with chromosome segregation at mitosis [10]. This type of treatment is usually local and targets specific areas either by high energy irradiation from external source or by the implantation of radioactive sources directly into or near the tumor. Both normal cells and cancer cells are affected indistinctively. However, normal cells treated with radiation arrest their cell cycle until they have repaired the damage to their DNA, thanks to the cell-cycle checkpoint responses. Cancer cells instead generally have defects in their checkpoint responses and they may continue to divide after irradiation, only to die after a few days because of the unresolved genetic damage. Even though the cells in a tumor have evolved to be unusually tolerant to minor DNA damage, they are hypersensitive to the much greater amount of damage that can be created by radiation. According to this principle, the treatment would preferentially kill cancer cells rather than healthy ones.

For sake of completeness, it is worth mentioning that also many more therapeutic approaches exist, such as photodynamic therapies, targeted-therapy with small-molecule drugs or monoclonal antibodies, blood stem cell transplants, hyperthermia and hormone therapy.

Thanks to extensive and ongoing study on cancer, remarkable progress has been made towards a better understanding its development and behaviour, which has led to major advances also in its treatment. However, cancer remains one of the most critical challenge in clinics due to several reasons, such as inter- and intra-tumor heterogeneity, the mutations in many different genes contributing to cancer and its tendency to evolve and progress over time accumulating new mutations and possibly developing resistance to therapeutic drugs [11]. At the same time, therapies themselves face important limitations due to their intrinsic toxicity or partial specificity: even if techniques such as chemotherapy and radiotherapy are specifically designed to be targeting preferentially tumoral cells, they still produce significant damages in surrounding tissues or are associated to strong toxicity for some organs, producing both short term and long term of different grade of severity, up to the point of being discarded in case of adverse patient response.

1.2.2. Current approaches for therapy monitoring and prediction

Once the treatment for the patient is selected, the monitoring of the tumor evolution and the evaluation of treatment efficacy are of utter importance; indeed, such feedback is fundamental to modulate the therapy in order to keep an optimal balance between pa-

tient's benefits and exposure to adverse effects; in the unfortunate case of a initial negative response to treatment, early awareness allows to readily make adjustments towards a personalized and more suitable approach. The currently available and established monitoring techniques include first of all exams based on imaging, such as coherent tomography (CT), positron emission tomography (PET) or magnetic resonance imaging (MRI). Thanks to these methods, it is possible to obtain images of the tumoral mass and to therefore evaluate the progression of cancer by the changes of its size. This approach however is limited by being unable to provide real time biochemical information of treatment response.

In the field of predicting therapy response, the use of experimental models such as cell lines or 3D models (e.g. organoids) to simulate the effects of therapies have been proposed to identify the therapy mechanism of action. However, the outcomes obtained are often linked to poor patient specificity, that can lead to discrepancies in the actual therapeutic response, thus hindering the reliability of the predictions and limiting success rates. The use of animals (i.e., xenografts) allows the reproduction of the tumor microenvironment (TME). The TME is a complex ecosystem surrounding tumor cells, composed of cancer cells, immune cells, and stromal cells (fibroblasts, endothelial cells, pericytes). It also includes non-cellular components such as the extracellular matrix, cytokines, chemokines, and growth factors. The TME is dynamic, with continuous interactions among these elements that influence tumor growth, metastasis, immune evasion, and therapy response. Therefore, it plays a key role in cancer development and evolution, suggesting its characterization may be significant for predicting the outcomes of a treatment. This is true, for example, for immunotherapies based on immune checkpoint inhibitors.
Current gold standard for tumor and TME evaluation is given by inspection of stained tumor slices (H&E) or by immunohistochemistry tecniques (IHC); both of them show however non negligible limitations. In particular, H&E images analysis cannot provide quantitative information and their interpretation remains significantly operator dependent; on the other hand, IHC techniques for ICI treatment still lacks of effective predictive biomarkers or thresholds. An example of this is the direct assessment of PD-L1 expression, which is not fully reliable since it can be transient, shows an intrapatient and intratumor heterogeneity, is limited by the usage of different thresholds for PD-L1 positivity and does not take into account factors that could impede the anti-PD1 or anti-PDL1 therapy response. Baseline tumor-infiltrating lymphocyte status instead cannot be provided with an absolute cutoff that could enable its use as predictive biomarker, while results based on association of neoantigens with immunotherapy benefit are strongly patient specific and cannot be generalized. All these considerations and a more extensive treatment can be found in [12].

Regarding the monitoring of the efficacy of treatments, one major problem is that, in most of cases, tests have to be performed at an advanced stage of the therapeutic course. Therefore, before having insights on whether the treatment is producing a positive response, patients must undergo therapy for a significant amount of time, with the risk of also being subject to negative health side-effects due to the intrinsic toxicity of many approaches. For example, CT and MRI for monitoring radiation therapy effects are to be used 2–3 weeks after completion of therapy, while other techniques like (PET) of fluorodeoxyglucose (FDG) can measure functional tumor response with reliable results after 8–12 weeks [13].

Spending time on an ineffective therapy is an issue since the cancer treatment is somehow a 'race against time': in the same way as for the screening, the earliest the tumor is identified and cured, the better it is, while if time passes and the tumor has time to grow and develop, the chances for a positive prognosis significantly reduces. In addition, in several cases malignant tumor can metastasize, making treatments much more complex. Finally, on one hand, prior insights on therapy efficacy can be determinant in the clinical decision-making process, while on the other, early warnings on therapy ineffectiveness allow oncologists and healthcare professionals to modify or fine-tune the therapy accordingly towards a more patient-specific solution.

In this overall context, a wide clinical need exists for an approach that is omics-like, fingerprint-based, label-free, nondestructive, rapid, and noninvasive.
As it has been discussed in the previous paragraph, Raman spectroscopy could be a good candidate technique that may be able satisfy these main requirements.
The use of Raman spectroscopy in clinical and medical field is not new; for example, Raman spectroscopy is being already successfully used for tissue diagnostics and cytopathology, and in the specific field of oncology it has already proved to be able to differentiate between tumorous and nontumorous tissues [14]. In the context of therapy effects characterization, Raman spectroscopy proved to be suitable for quantitative assessment of the molecular composition of lung and HN tumors xenografts subjected to radiotherapy and to probe radiation-induced alterations [13], while regarding the field of immunotherapy, in [15] responses of CT26 murine colorectal tumor xenografts to ICI treatment of anti–CTLA4 and anti–PD-L1 antibodies have been compared. Trials for in vivo studies have been made with the same cell lines of [13] in [16].

1.2.3. Case study: HNSCC

Head and neck squamous cell carcinoma (HNSCC) constitutes the vast majority of the broader category of head and neck tumors; it develops from the mucosal epithelium in the oral cavity, pharynx, larynx and sinonasal tract and is classified as the sixth most common cancer worldwide, with estimates foreseeing an increasing trend in incidence and occurrences in the upcoming years [17]. HNSCCs are generally associated with tobacco consumption and alcohol abuse; furthermore, for the specific case of pharynx cancers, it has been recognized a link with infection with oncogenic strains of human papillomavirus (HPV), especially HPV-16. Since this distinction is not limited to the etiology of the disease but extends to gene expression, and to mutational and immune profiles, HNSCCs are classified in the two separate categories of HPV-negative and HPV-positive HNSCC [18].

HNSCC tumors are mostly diagnosed in adult and elder population (median age of around 66 years) and at a late-stage of cancer development, due to the commonly asymptomatic nature. The principal modalities that are employed to tackle the disease include surgical resection of the tumoral tissue, radiation therapy and immunotherapy. Depending on the location and extension of the tumor, one treatment may be preferred over the others, but in many cases multimodal approaches are required, with a combination of the mentioned interventions and chemotherapy.

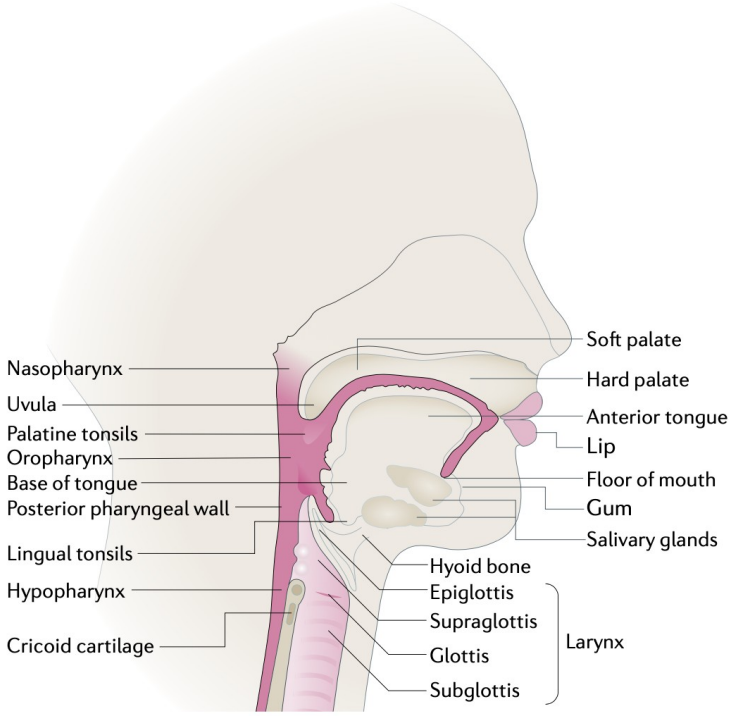


Figure 1.5: **Locations of Head and Neck tumors [17]**.

The challenges and needs described in the previous paragraph are relevant also for the specific cancer category of HNSCC.

Raman spectroscopy has been already successfully applied to study HNSCC, both for ex vivo and in vivo investigations, starting from tissue characterization ([19], [20]). By making use of classifications models such as PCA-LDA or PLSDA, some works have proven the capability of differentiating between tumoral tissue and healthy tissue ([21], [22]), or between healthy tissue, oral squamous cell carcinoma (OSCC), moderate and severe heterogeneous dysplasia (the latter being an abnormal concentration of cells which is not considered a tumor but may develop and become one) [23]. Other studies show the potentiality of surface enhances Raman spectroscopy (SERS) technique, which is emerging as a tool for evaluating the sensitivity of HNSCC patients to neoadjuvant immunotherapy by serum analysis [24] as well as for a broader range of applications [25]. Up to date, however, the application of Raman spectroscopy on HNSCC samples is relatively limited, and the small number of samples studied is a critical aspect affecting the accuracy of the corresponding results [26]. In addition, a very limited number of studies can be found in literature with the specific aim of characterizing and exploring the effects produced by anti-cancer therapies on head and neck tumors: in [16] the early effects of single-dose radiotherapy (within 48h after radiation exposure) on in vivo HNSCC xenografts of known

radiation sensitivity are investigated.

In this thesis, spontaneous Raman spectroscopy is applied on patient-derived tissue slices aiming at ex-vivo characterization of TME, cancer progression and therapy effects across a variable time range (between 0 to 5-7 days), with the benefits of a label-free, non destructive and chemically informative technique with respect to the currently standard methodologies. The long-term goal is to ultimately unravel possible resistance biomarkers that could allow the prediction of the tumor response to specific treatments, proceeding towards a model for tumor classification of therapy-resistant and non-resistant tumors.
This thesis work is performed at a project stage where investigations are focusing specifically on immunotherapy and chemotherapy, with provided samples subject to one of three different kinds of treatment:

- Regarding **immunotherapy**, the PDL1-PD1 pathway is targeted with an anti-PD-L1 treatment, as it happens for example in antibody-based, FDA-approved treatments with avelumab, durvalumab and atezolizumab [27]. In particular, an antibody that has a structure complementary to the one of PD-L1 receptors is introduced; in this way, the antibody can bind to the PD-L1 expressed on tumor cells, thus compromising the elusion mechanism previously described in and enhancing the capability of immune cells to attack malignant ones.
- For **chemotherapy**, instead, the effects of cisplatin treatment is studied, with two different dosages commonly used in clinical applications. Cisplatin is a platinum-based compound used as chemotherapeutic agent, which is able to form interstrand and intrastrand crosslinks that interfere with DNA duplication and repair, while also promoting the formation of reactive oxygen species which can eventually activate cell apoptosis, necrosis and autophagy [28].
- **PD-L1 treatment** analyses are also being carried out, which consists in providing tissue slices with the PD-L1 protein itself. This procedure is not to be confounded with the previously described anti-PD-L1 treatment; in particular, this second approach targets PD1 receptors on T-cells stopping the immune system action. The aim of this is not to fight cancer, but to leverage the tumor response regardless the content of immune cells in the extracted tissue section, so to be able to separate the changes produced by the activation of the immune system from the ones actually consequent to the anti-cancer treatment.

While the comparison of the effects of immunotherapy and chemotherapy are not yet being investigated in this work, a characterization and comparison of the two control models via Raman spectroscopy is needed to consider their use in future experiments.

1.3. Extracting information from Raman spectra: preprocessing and data analysis tools

When considering the use of Raman spectroscopy in various fields, including biomedical applications, the challenges are numerous, from the design and construction of suitable devices and measurement setups, to the design of experiments and data acquisition, but these are not the solely ones. A significant part of the work consists in extracting the true Raman signal from the data collected, to be followed by the analysis and interpretation of the results in such a way that relevant and reliable biochemical, biological, and medical information can be derived.

From a general point of view, two main steps can be identified in the data analysis process: the first is commonly referred to as **data preprocessing**, which could be defined as a "preparation" of the datum itself that enables the subsequent second step, which is the **statistical analysis** on spectral data. More precisely, the second part aims at deriving from the the datum insightful physical or biochemical information, by means of tools ranging from statistics to the most recent deep learning and AI models.
This, however, does not imply that preprocessing is of minor importance with respect to the following statistical analyses; on the contrary, this step is fundamental and indispensable. Firstly, preprocessing allows for the extraction of the information of interest by removing contributions from other phenomena, as it will be explained shortly. Additionally, data preprocessing is necessary to enable quantitative analysis by removing dependencies in the collected data from characteristics related to measurement parameters, such as the power of the incident radiation that originates the scattering process. In general, preprocessing aims at minimizing experimental variability and standardizing the data, thus facilitating repeatability of measurements and outcomes.

1.3.1. Preprocessing

Many examples of preprocessing of Raman spectral data can be found in literature; in general, some steps are shared among all of them and have been recognized as essential for a proper data pre-treatment, namely spike removal, instrument calibration, baseline correction, smoothing and baseline correction [29]. However, no pre-established, standardized and universal procedure exists: the optimal choices are tightly dependent on the specific implementation of the technique, the samples investigated, the case study, and the goals of the latter [30]. From this perspective, designing a robust pipeline suitable for a large dataset is far from trivial. Preprocessing steps that differ in terms of included

and excluded steps, parameter values assigned to the selected algorithms, and even the order in which the steps are performed can lead to different results at the end of the entire process.

In this context, one can consider for example the differences that might arise when the data being processed corresponds to individual Raman spectra, as opposed to hyperspectral maps. As mentioned earlier, Raman spectroscopy also allows for highly informative imaging of the sample, where each pixel in the image corresponds to a Raman signal collected at an exact location. The data may consist of a complete spectrum in the case of spontaneous Raman spectroscopy, or individual spectral features when using coherent techniques. By employing this acquisition modality, it is possible to combine and correlate spatial-morphological information with biochemical composition, that is actually what makes the choice for this kind of approach particularly appealing in many studies. Most preprocessing tasks for Raman maps are similar to those required for individual Raman spectra, but at the same time new challenges arise, such as distinguishing different regions of interest within the hyperspectral image. A clear example is the need of separating the pixels that correspond to the biological sample from those that do not contain any portion of it, which instead are entirely located in the substrate on which the sample is deposited (microscope slide, petri dish, cuvette, etc.).
As Raman maps include spatial information, preprocessing is expanded with tools that make use of this information to treat the data, enabling, for example, to enhance local uniformity between adjacent pixels, assuming that, at a local level, the sample is characterized by a reasonably small spatial variability. This can be the case when the specimen is sampled by small size steps comparing to the dimensions of the relevant structures characterizing it.

With that in mind, a core principle not to be overlooked is to avoid introducing any kind of bias into the data. A well-designed pipeline should in fact extract the Raman scattering information with the minimum possible alteration. For example, spectral smoothing algorithms are designed to remove noise terms modulating the real Raman signal, but attention has to be paid not to eliminate or deform peaks that are of biochemical relevance; at the same time, when noise features are particularly strong, also the opposite problem may emerge, that is the appearance of peaks that are instead artifacts, mimicking those belonging to true biological signals.
Another critical aspect is found in the baseline subtraction step, since removing an overly aggressive baseline may deplete Raman peaks intensities, producing a final result which does not anymore truthfully represent the original one.
For this reason, a careful monitoring on the effects of every single passage is important;

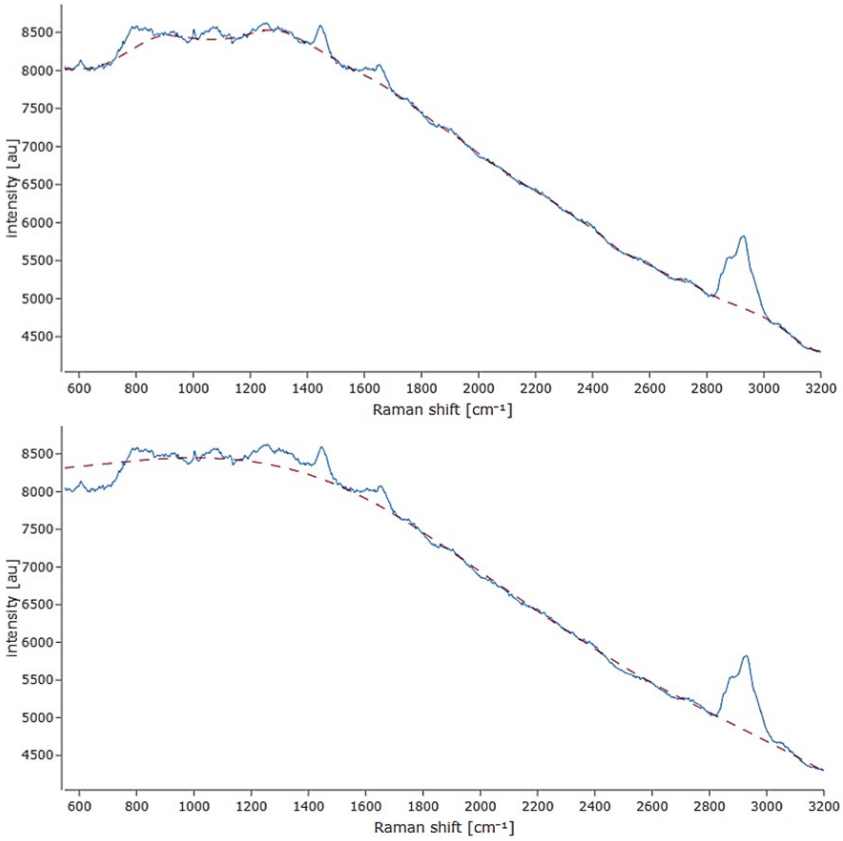


Figure 1.6: **Two examples of improper baseline correction**: above, the baseline is too aggressive and removes a portion of the fingerprint Raman peaks intensity; below, the baseline is too rough and cuts through the peaks at low wavenumbers. Images produced with *Ramapp* [31].

this, however, becomes particularly tricky with large datasets, which is also the case for the study that is carried on in this work. In particular, one can reasonably understand that inspections at single spectrum level is unfeasible when working with a quite extensive set of Raman maps, each of them comprising an elevated number of individual spectra; when a preprocessing pipeline of this kind is to be designed, compromises must be found to ensure good performances that are shared throughout the whole dataset.

To summarize the points discussed, a significant challenge that is also a part of this work consists in building a process that can be applied to a large and complex dataset (more than 300, 120x120 Raman hyperspectral maps), with fixed steps and parameter values for all spectra, allowing for effective and consistent batch preprocessing in preparation for statistical analysis.

1.3.2. Data analysis

Moving now to the interpretation of Raman data, one of the most immediate and straightforward approaches is to identify the position along the Raman shift axis of the observed peaks, in order to retrieve information on which biochemical components present in the sample being measured. By comparing with literature information, each Raman peak can be assigned to its corresponding molecular vibration. Sometimes, in the case of biological applications, the peaks is directly associated to the metabolites where the corresponding bond and vibrational mode can be found. For example, the Raman peak at around 793 cm^{-1} is often referred to as DNA since it is originating from C'-O-P-O-C'3 phosphodiester [32] that constitutes its structural backbone. By associating each peak that is visible in a Raman spectrum to the corresponding molecular oscillation/biochemical compound, one can retrieve the composition of the specimen that is subject to the measurement.

Univariate analysis

Going beyond the local chemical characterization in the position where the Raman spectrum has been collected, a wide range of more complex analyses can be carried out. Univariate analysis follows a simpler approach, focusing on specific characteristics by examining individual spectral features. Examples of univariate analysis include calculating and comparing the peak heights corresponding to the analytes of interest, especially when it can be reasonably assumed that there is a direct correlation between peak intensity and analyte concentration.
Another informative parameter is the width of the Raman peaks, which can be associated with the crystalline purity of the sample; this property is of particular interest in many fields and is often investigated, for example, in studies related to materials science [33]. In particular, in the ideal case where only a specific bond or component is excited by the incident radiation, a very sharp and narrow peak would be observed at the corresponding wavenumber, resembling a delta function. In reality, other compounds are often located near the excited molecule (for example, point defects, impurities, or vacancies in crystalline structures), which influence its vibrational mode, causing a dispersion in the frequency of oscillation and thus broadening the peak around its central wavelength; the final resulting shape can be approximated as a Lorentzian curve.
Finally, another phenomenon of interest is the shift in the wavenumber of a known peak, which can be caused, for example, by mechanical stress or strain on the sample, changes in the chemical environment, or structural transformation.

Another kind of variable often investigated in univariate analysis is band ratios, which

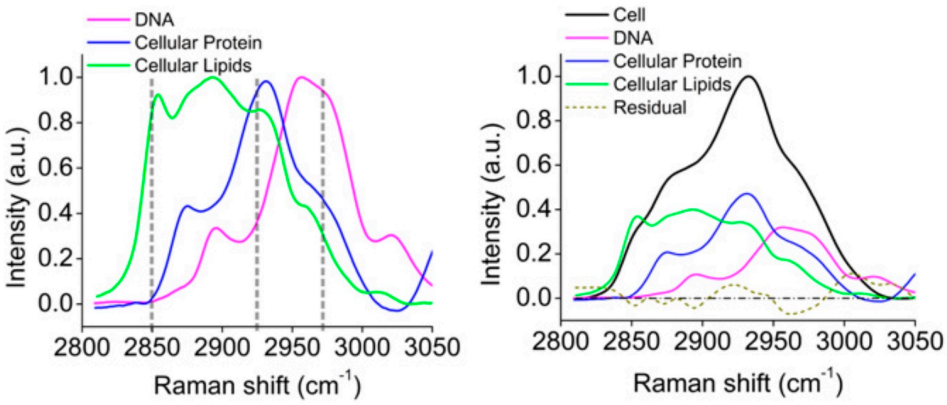


Figure 1.7: **Lipid, protein and DNA contributions to CH region peak**: on the left, the shape of the different contributions provided by lipid, protein and DNA to CH main peak. On the right, the final resulting peak and residuals after linear fitting of the components shown on the left. From [35].

are particularly useful in biomedical applications where the relative concentrations of different molecular components can provide diagnostic information. An example is the lipid-to-protein ratio, a widely used parameter that has been employed in many studies [34].

Generally, univariate analysis is particularly effective when peaks are clearly defined and separated one with respect of the other. However, especially in the biomedical field, a more common scenario is that spectral data tend to be more complex: overlapping peaks are often present and univariate approaches may fail if not aided with proper deconvolution and unmixing algorithms. A very common example for this is the CH region peak, which is produced by the convolution of different contributions mostly from proteins, lipids and DNA; the final shape depends on the relative concentration of each component (fig. 1.7).

Multivariate analysis

Since it is not easy to fully unravel many patterns of more complex nature by means of univariate analysis, other tools are also being used. In this context, a cornerstone technique is Principal Component Analysis (PCA), which decomposes spectral data into orthogonal components (principal components) that capture the maximum variance in the dataset. Typically, the first few principal components contain the most significant spectral information, while higher-order components tend to represent mostly noise and express less variance of the dataset. The scope of PCA is twofold: from one side it highlights the wavenumbers contributing the most to spectral variation, enabling the identification

of important molecular features that can be used for data separation, while on the other it allows for dimensionality reduction in the dataset. Dimensionality reduction based on PCA consists of representing Raman spectra with a set of principal components and associated scores; the more similar a single Raman spectrum is to the principal component, the higher the score.

A different set of approaches is represented by clustering techniques, such as K-means clustering, which aims to separate data into groups depending on their similarities. This method firstly selects a random set of values which are used as centroid location for the first iteration; all the elements on the dataset are then grouped in clusters depending on their distance to the centroids: every point is assigned to the cluster for which the distance is minimized. New centroids are then calculated by performing the average of the generated clusters, and the points of the dataset are re-assigned again; the process is iterated for an established number of times, or when no element is being classified on a different cluster with respect to the previous iteration.
K-means clustering is often used in vibrational spectroscopy an imaging tool, and has been employed on studies investigating of a wide range biological specimen, including for example tissue sections, cells, and in the analysis of human skin [36].

The dimensionality reduction and clustering algorithms described are just two possibilities among a larger category of methods, which are referred to as "unsupervised methods", since they work on data without any labels or classifications, separating them without prior knowledge of their original division. In contrast to them, supervised learning algorithms operate by receiving indications on how the data are classified, and build a model to make predictions and classification on new data provided. An example of such methods is Linear Discriminant Analysis (LDA), which operates similarly to PCA, but instead of finding orthogonal components that maximize variance across the entire dataset, it selects components that separate the categories in which each element is already classified. Supervised methods require two steps: a training step and a subsequent validation step; the performances of the algorithm are evaluated on how correct are the predictions made on the validation set of data.

Finally, it is worth mentioning that also new approaches involving the introduction of deep learning and AI algorithms are becoming widely used in recent years, leading towards automated, high-throughput, fast, and highly accurate processes with benefits in preprocessing, feature extraction, classification, regression and quantitative analysis [37]. At the current state of the art, the main limitations for these tools are mostly related to the scarcity of data to be provided to the model, which is instead essential for proper and effective training of the algorithms.In addition, deep-learning based approaches, such

as the implementation of CNNs, may provide interesting results in terms of prediction accuracy; however, these approaches are "black-box", meaning that it is not possible to access to the features and information mostly exploited by the algorithm, preventing reliable biochemical interpretation of their output.
Being the work presented in this thesis an early-stage contribution to a larger project, deep learning and AI use is being planned for the future activities but is not yet included at the present time in the analyses. For the same reason, no further or more detailed treatment will be provided in this chapter, since an extensive discussion would be required in order to be exhaustive on the topic, which would be outside the scope of this thesis work.

2 | Aim of the study

This thesis is part of the RAMPART (Label-Free Raman Approaches for Patient-Specific Biochemical Tumor Mapping to Monitor Personalized Anticancer Resistance Simulation) project, which poses as primary aim to develop a label-free, nondestructive and rapid methodology using Raman and multimodal imaging to identify biomarkers of tumor resistance, by monitoring the biochemical responses of ex vivo patient-derived tumor slices to both immunotherapy and radiotherapy.
Specifically, the objective of this thesis is to establish a methodology for acquiring Spontaneous Raman hyperspectral maps of HNSCC samples, develop a preprocessing pipeline, and perform a subsequent data analysis to evaluate and validate the previous steps. Beyond these objectives, the thesis will also allow to gain awareness of technical challenges, while at the same time uncovering the potentialities of the technique.

In detail, the work will focus on the following aims:

- Acquisition of Raman hyperspectral maps of HNSCC tissue slices, identifying the relevant regions of interest and establishing appropriate acquisition parameters.
- Design, test and optimization of a preprocessing pipeline, with particular attention on its scalability over the entire dataset of Raman hyperspectral maps.
- Development of Python tools to address the specific needs of the case study, involving in particular:
  - substrate identification and background signal subtraction;
  - artifacts identification and removal from the analyses;
  - reproduction of the tissue areas classification provided by histopatologists.
- Analysis of processed data through univariate and multivariate analysis, to gain insights on local biochemical features characterizing the samples, with the challenge of the implementation of algorithms allowing analyses across multiple maps at the same time.

3 | Materials and methods

3.1. Experimental setup

The experimental setup employed to perform the measurements is an home-built optical system for microspectroscopy, that allows measurements of spontaneous Raman scattering and spontaneous Brillouin scattering signals [38]. A schematic representation of the setup is provided in fig. 3.1.

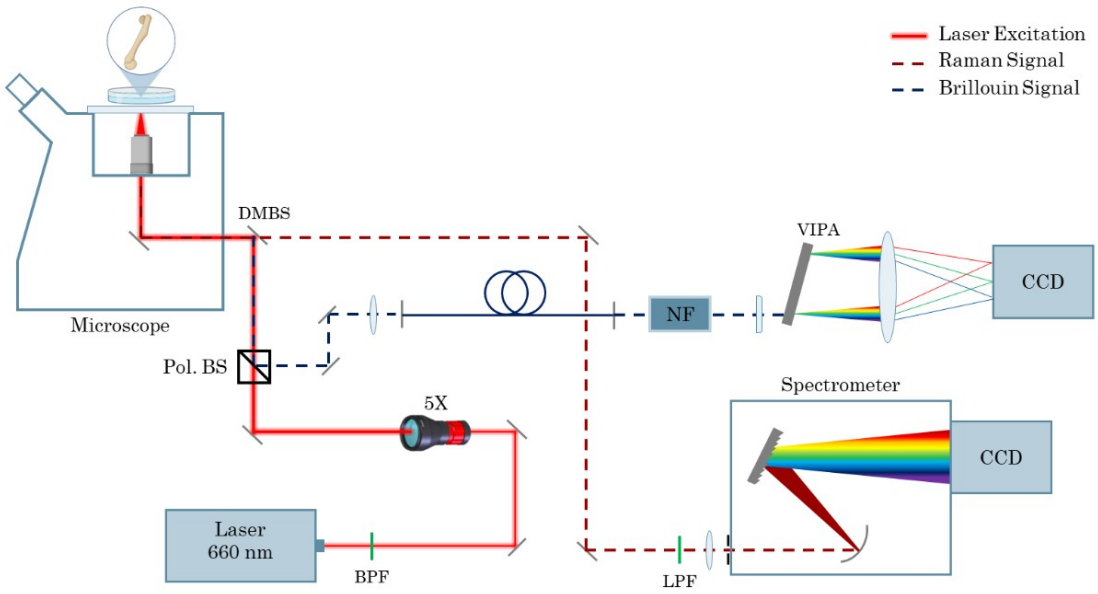


Figure 3.1: **Schematic representation of the experimental setup.**

The laser source used for excitation is the Cobolt Flamenco^{TM}, a continuous wave, diode pumped laser operating at 660nm with a maximum output power of 300mW. The excitation light is directed towards a shortpass filter to remove possible spurious peaks and then passes through a neutral filter, that can be chosen from a set of filters of different optical density by means of a rotating support. Their use allows to regulate the power of the excitation light impinging on the sample: since we are dealing with biological specimens, this

is very important in order to avoid damages such as burning. By making use of a beam expander, the spot size of the laser light can be tuned from the initial value (aperture of 700 ± 50 μm) up to a 5 times larger one, to better fit the back-aperture of the objective in use. In particular, the commercial microscope that is included in the setup is provided with four objectives with different magnification factors (5x, 20x, 50x and 60x); for the measurements presented in this work, Raman spectra are always collected with the 50x objective, whose specifications are summarized in table 3.1, together with the theoretical spatial resolutions that can be achieved.

A Semrock Brightline Di03-R660 single-edge dichroic beamsplitter (corner wavelength $\nu_c = 660nm$) has the dual function of reflecting the excitation light towards the microscope entrance and separating the back-scattered photons belonging to the Raman scattering from the Rayleigh one. Indeed, the latter is at the same frequency of the excitation radiation and is characterized by an intensity several orders of magnitude higher than the Raman one; its collection would be problematic since it would cause a saturation of the signal collected at the detector and obscure the much smaller Raman signal.

An Olympus IX73 confocal inverted microscope allows the acquisition of brightfield images, along with the collection of Raman and Brillouin spectra. For the first one, the light source and the condenser lens are placed above the specimen stage, while for the latter the laser light illuminates the sample from below, with the scattered light being collected on the same side. The microscope is equipped with a 2-axis motorized stage that enables both manual and electronically-controlled movement of the sample, allowing the selection of the excitation spot on the sample and the region of interested for 2D Raman hyperspectral maps. A two-staged turret provided with mirrors redirects light either towards a camera (when collecting brightfield images) or back to the optical path. Due to the frequency shift, the Raman signal now can be transmitted through the dichroic beam splitter and is collimated with a lens of focal length f=50mm inside the spectrometer (Princeton®IsoPlane 160), which is provided with a grating of 600 gr/mm groove density that spatially separates the signal components according to their frequency. Finally, the dispersed light impinges on a front-illuminated camera (PIXIS 256E, Princeton Instruments) featuring a sensor of 1024×256 pixels of size of 26μm$\times 26\mu$m and a 100% fill factor. An adjustable entrance slit, ranging from 10 μm up to 200 μm, controls how much Raman scattered light enters and regulates the spectral resolution of the system. Before reaching the spectrometer, light passes through one last filter, namely the 664 nm Semrock® RazorEdge LP02-664RU ultrasteep long-pass edge filter, to ensure the removal of any residual signal deriving from Rayleigh scattering.

Specification	Value
Magnification	50x
Numerical aperture	0.8
Working distance	1 mm
Immersion medium	dry/air
Lateral resolution	611 nm
Axial resolution	2.31 μm

Table 3.1: **Specifications of the objective employed for measurements (MPLFLN50X)**. For the calculation of spatial resolutions an emission wavelength of $\lambda_{em} = 740$ nm (1600 cm^{-1}) has been assumed [38].

In case of Brillouin scattering measurements, the excitation path from the laser source to the microscope is identical to the one previously described, while the collection path is different. Indeed, Brillouin scattered photons provide information at frequencies at the order of GHz, therefore with a frequency shift significantly lower with respect to the Raman scattering one, in the order of 10 cm^{-1}. The Brillouin signal is reflected by the dichroic beam splitter, then deflected by a polarized beam splitter and finally coupled into an optical fiber. The light coming out from the other end of the optical fiber passes through a common-path birefringence-induced phase delay (BIPD) filter, used to reduce background signal, and then directed towards a single-staged virtually-imaged-phased-arrays (VIPA) spectrometer. The signal is finally measured by the same CCD camera used for spontaneous Raman measurements. In the present work, no measurements of Brillouin scattering signal were performed; for this reason, the description of the Brillouin configuration will be limited to this brief introduction, as a complete and detailed treatment would exceed the scope of this work.

3.2. Samples description

The samples analyzed in this work are collected from patients of Jena University Hospital during surgical procedure for tumor excision, with the approval from the local ethical committee and in line with the declaration of Helsinki guidelines for research on human subjects by the World Medical Association [39].
The dataset used in this thesis includes samples from 5 patients, which will be referred to with their respective codes: SC7, SC8, SC15, SC17, SC23. The steps involved in sample preparation are schematically summarized in figure 3.2. Tumoral masses extracted from patients are preserved in RPMI 1640 (1% PenStrep, 1% Nystatin), cut with a chopper in 350µm-thick slices and cultured in a 6-well plate for a certain amount of time, up to a maximum of 6 days; one part of them undergoes a certain type of treatment (PD-L1 protein treatment, anti-PD-L1 treatment or chemotherapy with cisplatin) starting from day 1, while the others serve as a reference for therapy monitoring. All the 350µm-thick slices are then frozen and preserved at -80°C. Prior to shipment, slices are cut again in thinner sections (12 µm) with a cryostat, fixed with neutral buffered formalin (10%), washed two times with MilliQ water, vacuum dried and then put on 1mm-thick quartz slides. Slices are finally shipped to Politecnico of Milano, where the Raman measurements illustrated in this work are performed.

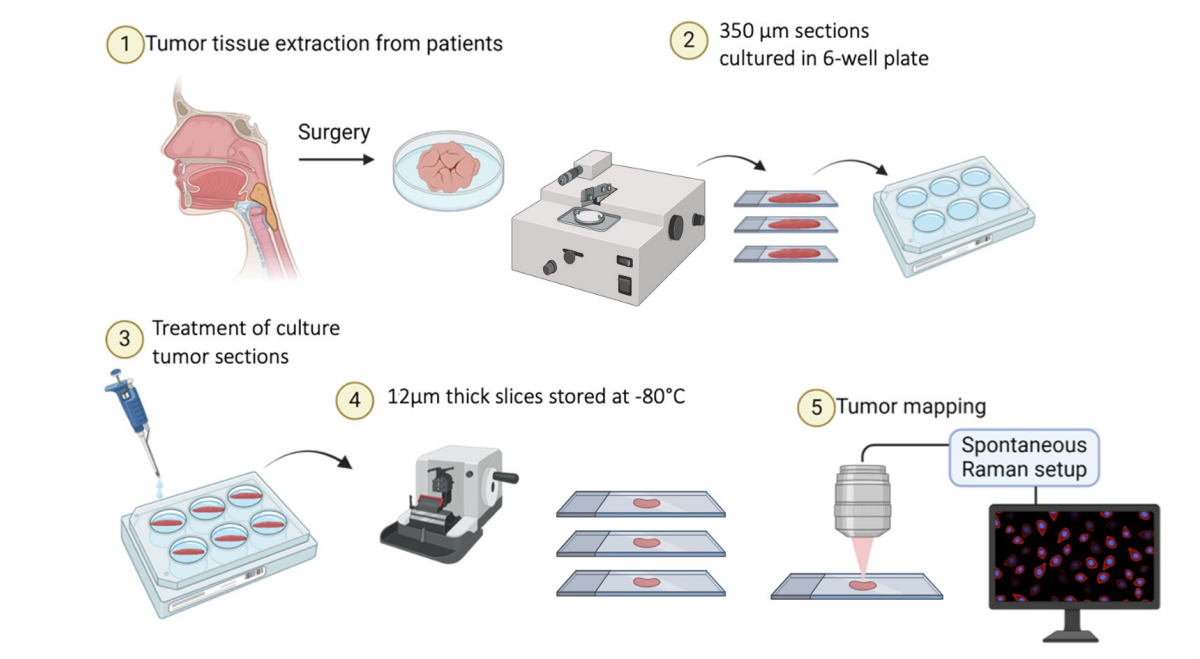


Figure 3.2: **Steps of the sample preparation procedure**. Image produced with *Biorender*.

The PD-L1 protein treatment for patients SC7, SC8, SC15, SC17 is performed by providing human PD-L1 (Advanced BioMatrix 5126-0.1MG), with a final concentration of 10 μg/mL, performing a change of medium and introduction of new PD-L1 on a daily basis. At the moment of the present work, only for patient SC23 an immunotherapy and chemotherapy simulation has been performed. In particular, cisplatin from Merck 232120-50MG is used as chemotherapeutic agent with two different concentrations: 3,33 μM and 6.66 μM. For anti-PD-L1 immunotherapy treatment, Anti-Hu CD274 [MIH1] Invitrogen 14-5983-82 with 10 μg/mL is used. Also in this case, as for PD-L1 protein treatment, every day the sample medium is changed and the treatment is applied.

For each tissue slice that is intended to be measured, one contiguous 12µm tissue slice is saved at Jena University Hospital, where the hematoxylin and eosin (H&E) staining is applied. Thanks to this procedure, which is widely used as a gold standard for histology and medical diagnosis, different structures within the tissue appear with different shades of purple and pink. The images of these stained slices are provided to us, together with annotations by trained histopathologists, indicating whether a specific area of the tissue slice is considered to be tumoral or healthy (either generically or with details on the kind of healthy tissue that is being observed).

Measurements are performed for every timepoint available, treatment condition and area classification (tumor tissue and healthy tissue). In order to identify correctly the regions of interest to be measured, brightfield images of the provided tissue slices are collected with the x5 objective and are compared to the H&E stainings and annotations (figure 3.3. Each map is intended to cover a single type of tissue on the slice; in other words, each single map is classified either as healthy or tumoral. As it will be explained later more in detail, this procedure is not always possible and sometimes it is unavoidable to collect Raman maps of heterogeneous composition, including a both regions classified as tumoral and healthy. In this cases a more refined intra-map, pixel-wise classification is performed by means of a custom Python code that allows the manual annotation and labeling on Raman maps.
Following the acquisition of every map, brightfield images are acquired with the x20 objective, precisely including the same area covered by the hyperspectral image, thus providing a reference especially for the separation between substrate and biological specimen, but also in case the abovementioned manual annotation of the Raman map is needed.

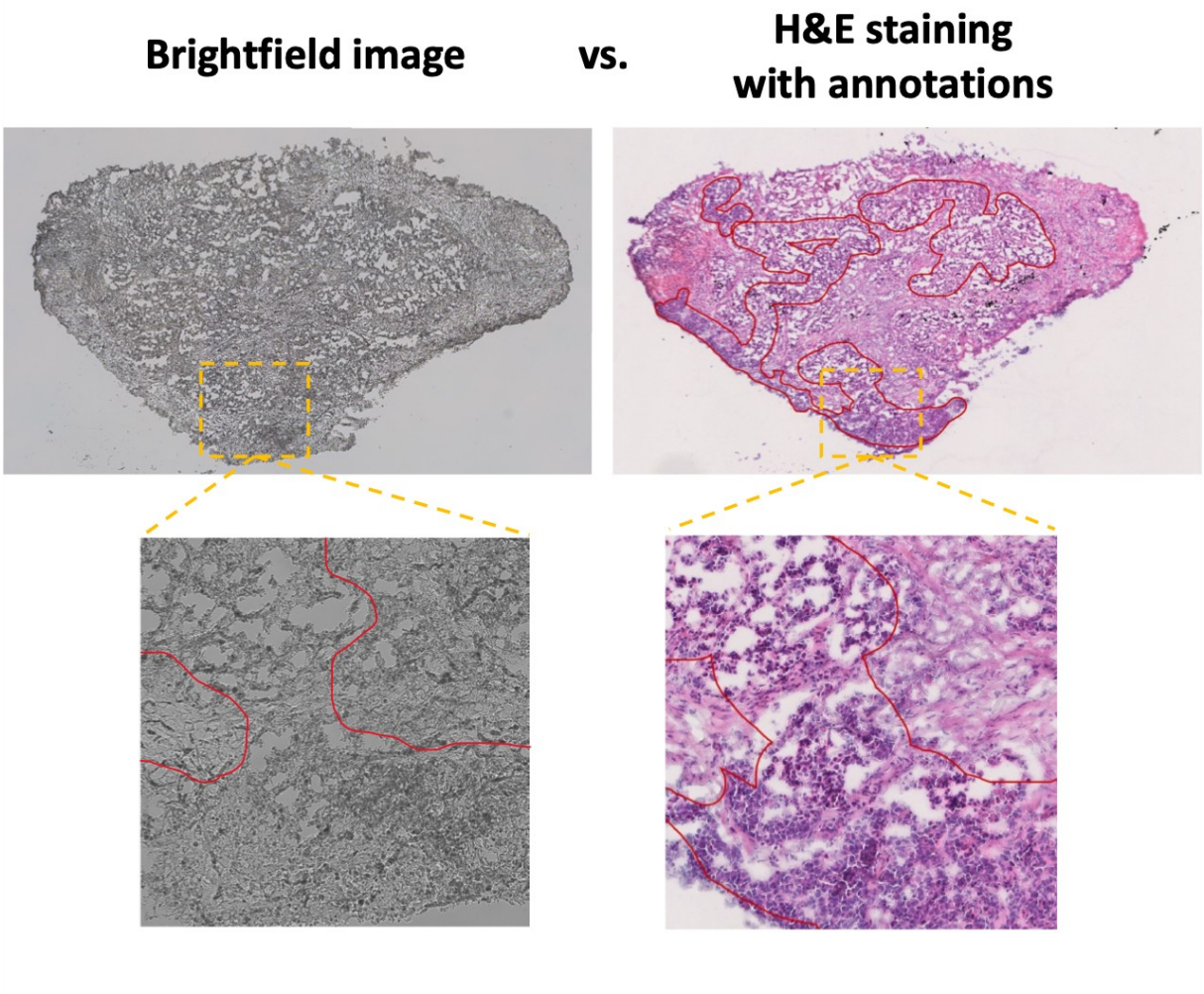


Figure 3.3: **Procedure to select regions to be measured**: the brightfield image of the tissue slice (left) is compared to a contiguous slice that is stained with hematoxilin and eosin (H&E staining) together with annotations by histopatholigists (right). The area inside the red line in H&E stained images is classified as tumoral. In the corresponding brighfield area, a tentative reproduction of the annotations is drawn.

The number of tissue slices available for measurement is 128; in particular, two replicates of slices belonging to patient SC15 have been provided, while for patient SC17 the replicates amount to six. Details on available timepoints and treatments are summarized in tables 3.2a and 3.2b, while table 3.2c contains the calculation of the total number of tissue slices. In particular, for every patient the following slices are being provided:

- one slice at timepoint t0, which corresponds to a biopsy of the tumor: the slice is not subject to any treatment or cultivation after the extraction;
- slices for N_{Time} different timepoints;

- for every timepoint, N_{Treat} treatment options are available.

The set of tissue slices indicated with these parameters constitute one replicate; the total number of slices per replicate amounts therefore to:

$$N_{1rep} = t_0 + N_{Time} * N_{Treat} \tag{3.1}$$

Since, for every patient, a certain number of replicates N_{Rep} is provided, the total number of slices to be measured for each patient is:

$$N_{tot} = (t_0 + N_{Time} * N_{Treat}) * N_{Rep} \tag{3.2}$$

Information on the final subsets selected for performing analyses can be found in paragraph 4.2, while indications on the nomenclature adopted for managing data are illustrated in Appendix A.

a)

Patient	t0	t1	t2	t3	t4	t5	t6	t7
SC7	✓	✗	✗	✓	✓	✗	✗	✗
SC8	✓	✗	✓	✗	✓	✗	✗	✗
SC15	✓	✓	✓	✓	✓	✓	✓	✓
SC17	✓	✓	✓	✓	✓	✓	✓	✗
SC23	✓	✓	✗	✓	✗	✓	✗	✗

b)

Patient	Control	PDL1	Anti-PDL1	Cisplatin 3,33 µM	Cisplatin 6,66 µM
SC7	✓	✓	✗	✗	✗
SC8	✓	✓	✗	✗	✗
SC15	✓	✓	✗	✗	✗
SC17	✓	✓	✗	✗	✗
SC23	✓	✗	✓	✓	✓

c)

Patient	SC7	SC8	SC15	SC17	SC23	
t0	1	1	1	1	1	
N_{Time}	2	2	7	6	3	
N_{Treat}	2	2	2	2	3	
	5	5	15	13	10	
N_{Rep}	1	1	2	6	1	
Total	5	5	30	78	10	128

Table 3.2: **Dataset information: available timepoints (a), treatments (b) and tissue slices number (c) for each patient**. Ticks and crosses indicate whether, for a specific patient, samples for a specific timepoint (table a) or treatment condition (table b) have been provided or not.

3.3. Data collection

3.3.1. Measurements and acquisition parameters

For every patient, timepoint, area and treatment condition, two different kind of measurements are performed:

- Single Raman spectra: a set of Raman spectra is collected in 30 randomly selected, non-coincident points falling within the region of interest of choice (one for the tumoral tissue and another for the healthy). This dataset is meant to be used for a preliminary investigation of the sample, giving up in first instance the spatial-morphological information in favor of an higher quality of the collected spectra. Indeed, parameters are chosen with the aim of improving as much as possible the signal to noise ratio, with looser constraints, for example, on the acquisition time and laser power employed. Under these considerations, the choice was finally to measure with an output laser power of 150mW and exposure time of 2 seconds.
- Raman hyperspectral maps: in order to characterize properly the samples, a larger number of spectra is reasonably required. In addition to that, maps of extensive regions can also provide insightful information from the spatial-morphological features of tissues. These are the reasons that led to the choice of basing the core of the data analysis and study on Raman maps instead of single spectra. The acquired maps consist in 121x121 pixels hyperspectral images, for which each pixels correspond to a full Raman spectrum. Each point of the pixel correspond to a 5µm step in the sample plane, which results in an overall field of view of 605x605μm^2 area. For a more conservative approach in terms of risks of sample damaging, the laser power was set to 120mW, while the exposure time had to be reduced to 500ms as a compromise between quality of collected data and time required for performing a full-map measurement. With the listed parameters, the total time required for a single map acquisition is around 170 minutes. Spectra are acquired with a vertical binning of the CCD averaging across the 256 vertical pixel, resulting in 1024 horizontal pixel (consequently, Raman shift values) measurements.

Single Raman spectra measurements are performed manually to ensure a proper selection of the excitation spot, with particular attention to the optimal z-axis focus (depth inside the sample), while Raman maps are automatically collected thanks to the automated electronic control of the translating stage by means of Micromanager software and Python codes making use of related instrument-control APIs.
In the second case, since at present time no auto-focus function is available for the setup,

the position of the laser focus along the z direction has to be fixed for the entire map; the final choice is, for this reason, a compromise that allows to get a good Raman signal from the sample while limiting the penetration inside the quartz substrate.

The analysis of single Raman spectra has been performed aiming at preliminary investigation and for testing the performances of some preprocessing steps, such as background signal subtraction with the EMSC algorithm, characterized by a rescaling of the substrate-related signal to each sample Raman spectrum (details are provided in paragraph 4.1.2). For this reason, results produced with this dataset are not reported in the following paragraph 4.2 dedicated to experimental results.

3.3.2. Wavenumber and intensity calibration

An essential step that must be performed in order to correctly interpret the collected data is the calibration of the system, both in terms of frequency and intensity. In particular, what we are able to obtain from measurements consists in the intensity measured by each horizontal pixel of the CCD (1024 values); however, no information about the wavenumber or Raman shift value for each pixel is available a priori. In order to express the Raman intensity as a function of a physical meaningful quantity (wavelength of the collected light or corresponding Raman shift value in wavenumber units) a calibration procedure has to be performed. In particular, to achieve this pixel-wavenumber conversion, a common method consists in measuring the Raman spectrum of some substances, usually preferring materials and chemical compounds which produce an intense Raman signal, with sharp peaks located at very precisely known wavenumbers. Also generic light sources with emission spectrum sharing these same characteristics can be employed. In this work, the choice of the calibrants goes to the following:

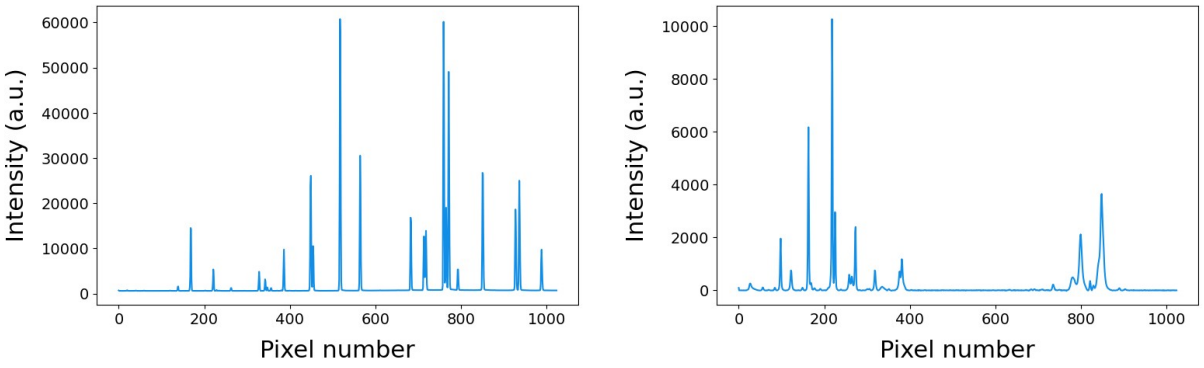
- Toluene, whose Raman signal is measured using the 50x objective, with an output laser power of 100 mW, an integration time of 100 ms and averaging for 5 frames;
- ArHg lamp (Avantes AvaLight-CAL-Mini), whose light is focused directly inside the spectrometer, with an integration time of 0 ms.

The spectrum of the ArHg lamp reported in the documentation and the theoretical Raman signal produced by toluene are compared with the ones obtained experimentally, so to assign the correct wavenumber value to each pixel. The use of two calibrants instead of a single one aims at adjusting the calibration of the wavenumber including the slight variability of the output wavelength of the laser: indeed, to move from x-axis values expressed in terms of wavenumbers $\tilde{\nu}$ to wavelength λ and vice-versa, the wavelength λ_0 of the impinging radiation giving rise to the Raman scattering process must be known:

$$\tilde{\nu}\ [\mathrm{cm}^{-1}] = \left(\frac{1}{\lambda_0[\mathrm{nm}]} - \frac{1}{\lambda[\mathrm{nm}]}\right) \times 10^7 \tag{3.3}$$

The calibration is therefore configured in a two-step process:

- Starting from the assumption of having an unknown excitation laser wavelength, its value is computed by means of the collected ArHg signal and using (cite formula). The pixel-wavelength conversion is performed too.
- By means of the collected toluene spectrum and its comparison with the theoretical one, pixel-wavenumber conversion is achieved.

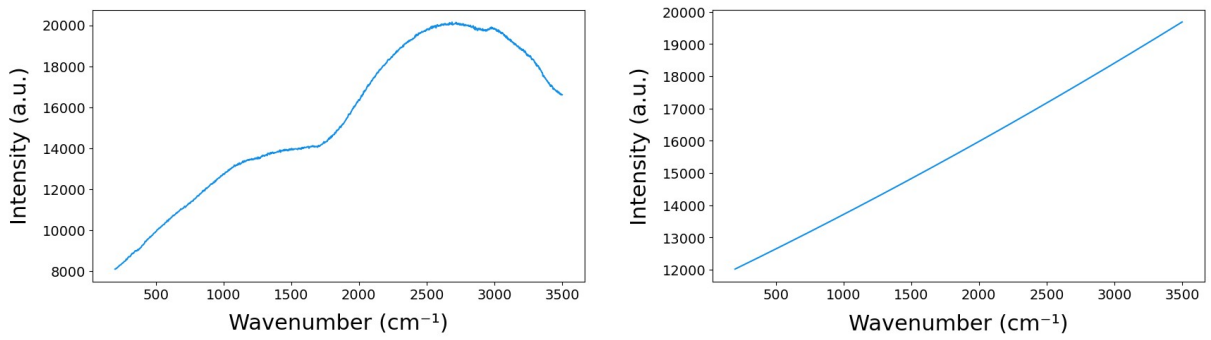


(a) (b)

Figure 3.4: **Measured Raman spectra of ArHg lamp (a) and toluene (b) used for wavenumber axis calibration**.

Another important calibration step consists in the intensity calibration, which is necessary because the CCD sensor has a frequency-dependent quantum efficiency and therefore weights differently each spectral components across the entire wavenumber range. In addition, this intensity response function is not always the same, but may change over time, and is also affected by many other external factors such as, for example, temperature, excitation wavelength and sample geometry. To make the analyses quantitatively meaningful, an intensity calibration is performed by collecting the light emitted by a white lamp and comparing it to its theoretical spectrum (see Fig. 3.5), computing a coefficient $q_e(\lambda)$ for each discrete value of λ in the Raman shift axis.

Finally, the last step that enables comparison of spectra with simple operations (sum, differences, band/peak ratios...) and ensures consistency in the preprocessing and analyses is the interpolation of the collected spectra with a specific set of wavenumber values, so to set a common x-axis to all the data measured. In our case, the choice fell to a set of 1024 equally separated values of Raman shifts ranging from 200 cm^{-1} to 3500cm^{-1}.



(a) (b)

Figure 3.5: **Raman spectrum of the white lamp: measured (a) and theoretical (b).**

All the listed data prepraration steps have been performed by means of already provided Python codes or via Ramapp (x-axis interpolation of Raman maps).

3.3.3. Data preprocessing

After the collection of Raman signals and the preliminary calibration and harmonization steps just illustrated, an accurate preprocessing of data has to be performed before proceeding with quantitative analyses, both in the cases of single Raman spectra and hyperspectral maps.

Some principles and good practices for Raman data preprocessing are going to be illustrated in this paragraph; however, it is important to keep in mind that no perfect nor universal pipeline can be established. Since the design of a proper procedure for this work has been one of the main focuses of this thesis, its detailed discussion will be held more in depth in 4.1.

First of all, one of the most important steps consists in the removal of cosmic rays signal, which affects any kind of Raman measurement system and cannot be removed *ab initio*. Cosmic rays refer to high-energy particles that can be collected during measurements, whose presence is noticeable inside the Raman spectrum acquired by the occurrence of spikes, that are very sharp and high intensity peaks that appear in random positions across the wavenumber axis. Their presence in the collected spectrum is detrimental for other preprocessing steps and can produce downstream negative effects on data classification and analysis. For this reason, Z-score algorithm is adopted to spot their presence and readily correct them before proceeding further.

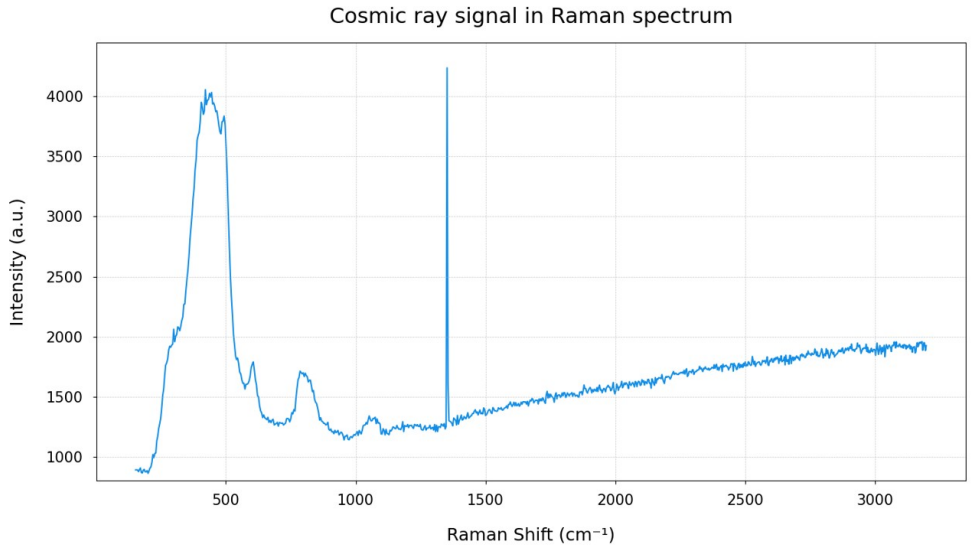
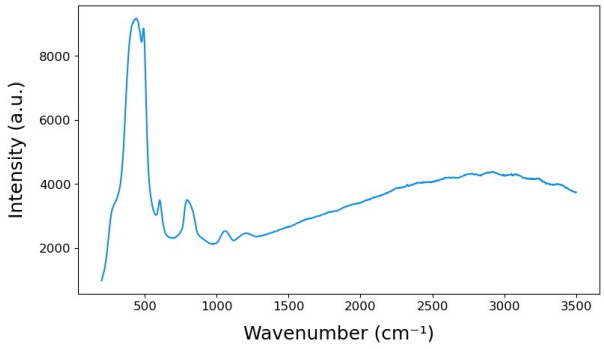


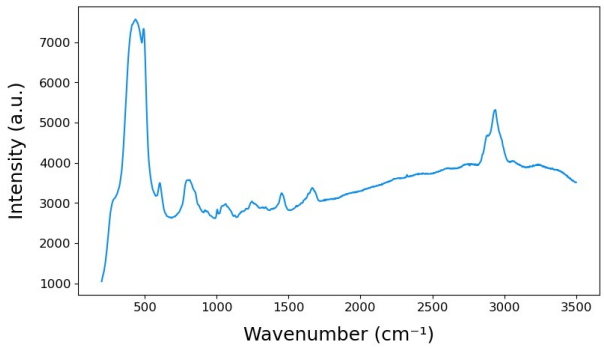
Figure 3.6: **Example of signal from cosmic rays**

A second but nonetheless relevant contribution not originated from the sample is the one produced by the substrate material onto which the specimen is deposited, such as petri-dishes and microscope slides, or other substances surrounding the sample, such as water or any other medium. In this work, in particular, tissue slices are provided on coverslips of 1 mm thickness made of quartz, which is a Raman-active material producing a signal that at low wavenumbers up to the middle of the fingerprint region (see fig. 3.7). This contribution is inevitably collected and is not trivial to be dealt with, since its features have different intensities depending on the acquisition conditions (thickness of the sample, position of the focus etc...). Two different approaches can be adopted:

- the substrate contribution is not removed and then is kept in consideration in the successive analyses;
- contaminants spectra are subtracted from the measured datum, with the challenge of optimizing the algorithm in such a way to avoid having residual substrate signal or creating artifacts due to an eventual improper choice of the rescaling factor.



(a)



(b)

Figure 3.7: **Raman spectrum measured inside quartz substrate**, without any sample (a) and Raman spectrum measured on the specimen (b), the latter featuring both peaks from biochemical components and peaks from quartz as in (a).

Successively, the computation and the subtraction of an appropriate baseline curve must be performed in order to better highlight the Raman signal features and remove the contribution from light that is being collected but is produced by other phenomena such as the autofluorescence of the sample. Also in this case, a wide set of methods is available for this purpose, each of them having better or worse performances depending on the shape of collected spectrum. In addition, it is fundamental to also carefully tune the specific parameters of the baseline estimation method of choice, since no deformations of the Raman spectrum must be produced: the baseline curve should smoothly reproduce the path of the collected signal but at the same time be such that there is no introduction of negative peaks or partial reduction of some Raman features.

Finally, since the intensity of the signal collected is dependent on many factors such as the excitation light power impinging on the sample, setup configuration, different substrate, local features of the specimen and many more, a normalization step is always to be performed. Min-max and L2 vector normalization, which are among the most common

and widely used methods, are those that will be employed in this work.

According to the aim of the study and possible prior knowledge on the analyzed samples, it is common practice to crop the Raman spectrum to remove wavenumber regions that are not considered of interest; for example, in the case of biological and biomedical applications, discarding the information contained in the silent region or at low Raman shift values. Removing non-informative spectral regions or restricting the analysis in a specific range of wavenumbers can indeed significantly improve the outcome of many other data manipulations and analysis.

The steps that have been illustrated up to now are very common and in a certain sense essential in all pipelines for Raman data preprocessing. All these steps are performed both for single spectra measurement and maps.
For the latter, however, an additional step has been introduced, which is the removal of the first row and first column of the collected map. Indeed, the actual dimensions of the collected map has been set to 121x121 pixels, to have a final 120x120 image covering a field of view of 600x600 µm after the removal. The reason behind this procedure is linked to the photobleaching effect affecting the autofluorescence light produced by the sample. In particular, as mentioned before, Raman spectra are superimposed to a broad baseline which is produced by other mechanisms, among which autofluorescence is the mostly contributing phenomenon. Autofluorescence signal intensity, however, shows a decreasing trend over time if the sample is kept under the excitation laser light: this process is referred to as photobleaching. The occurrence of this phenomenon is noticeable by comparing the first Raman spectra collected with the rest of the map: the first row of pixels is usually characterized by a larger average intensity because of the higher baseline deriving from autofluorescence. Passing then to the following rows, the excitation light will illuminate areas on the sample close enough to the previously illuminated points; the newly measured points will be therefore affected by photobleaching due to the previous illumination, and will present lower average signals. In order to avoid the presence of this difference affecting the first row and first column of measured points, the corresponding pixels have been discarded from each map (we will refer to this procedure as 'map crop')

The tailoring and establishment of the preprocessing pipeline has been performed by custom Python codes using the following libraries: *numpy, pandas, scipy, scikit-learn, ramanspy*. In the case of maps, also the open source web application *Ramapp*, developed by our group and designed specifically for visualizing, preprocessing and analyzing Raman hyperspectral images, has been used [31]. Finally, univariate and multivariate analyses have been entirely performed with custom Python codes and Jupyter notebooks, making use of the previously listed Python libraries and producing plots and graphs with

matplotlib and *seaborn*.

4 | Experimental results and discussion

4.1. Design and optimization of the preprocessing pipeline

The first part of the present work is focused on the establishment of a preprocessing pipeline to be applied on the entire dataset of the study. As mentioned in the previous paragraphs, both custom written Python codes and already existing tools (Ramapp) are being used for this purpose.
The results presented in this section are divided into two paragraphs; the first one is dedicated to the description of the preparation and the evaluation of a preprocessing pipeline exploiting the web app Ramapp and its source codes. The second one instead is focused on additional tools introduced by custom made codes to tackle problems of a different kind, which are specific for the study and go beyond the field of spectral data preprocessing.
The outcomes of univariate and multivariate analysis performed on preprocessed spectra will be illustrated later in paragraph 4.2.

4.1.1. Tested pipelines and comparison

The design of a proper preprocessing pipeline is based on the known principles and good practices illustrated in paragraph 3.3.3.

In order to test different pipelines and make some evaluations before adopting one of them for the entire dataset, a subset of 5 maps (one per patient) was chosen for testing purposes, with the criterion of including as many different areas, timepoints and treatment conditions as possible, aiming at generalizing the results and exploring the performances of the pipeline in measurements made on different tissue types. To simplify the process, the selected maps are such not to contain artifacts of any kind, such as strong autofluo-

rescence or burnt areas; in paragraph 4.1.2 will then be explained how the effectiveness of the pipeline can be preserved by introducing an automatic identification of pixels with artifacts, that allows to exclude them in the following steps of analysis. Indeed, as it will be illustrated, those pixels usually show a completely deteriorated spectrum which cannot be used anymore to extract biochemical information, regardless of the preprocessing pipeline applied. Including also those data in the statistical analyses would strongly influence the results obtained, yet at the same time discarding an entire map when their presence is limited to a very little number of pixels (as in most cases occurs) may be a too harsh approach, especially when considering clinically relevant regions of interest.

In particular, the maps included in the subset for testing are the following:

- 3b: patient SC8, day 2, PD-L1 protein treated, healthy area
- 6c: patient SC7, day 0, mixed area
- 24b: patient SC15, day 7, non treated tumoral area
- 37a: patient SC17, day 3, PD-L1 protein treated, tumoral area
- 54c: patient SC23, day 5, cisplatin treatment (3,33 μM)

Since the preprocessing was planned to be performed with Ramapp source codes, the preliminary testing of the different preprocessing steps was performed directly through Ramapp interactive interface. This approach allows for testing qualitatively the effects of different preprocessing steps, algorithms and parameters with the advantage of having a straightforward interface and rapid accessibility to the spectral data contained pixel by pixel. The main conclusions from this procedure regarding for the general structure of the preprocessing pipeline are described and motivated as follows. Each step will be corredated with sample images of the transformation of one pixel inside map 6c, to allow a visualization of the effects produced by the each pipeline step.

The first steps to be performed are the x-axis interpolation and the removal of the first row and first column of pixels of the collected map. As described in paragraph 3.3.3, these two operations can be regarded as a step that is preliminary even with respect to the real spectral preprocessing, and for this reason they are naturally introduced at the very beginning of the data manipulation and preparation.
Following this, the spectral axis is cropped between the wavenumber values of 650 cm^{-1} and 3050 cm^{-1}. This choice has been made considering that for wavenumbers lower than 650 cm^{-1} the signal coming from the quartz substrate is so intense that the Raman peaks belonging to that spectral region are completely overwhelmed. Even with a proper background signal removal, the residual signal at these specific wavenumbers cannot be

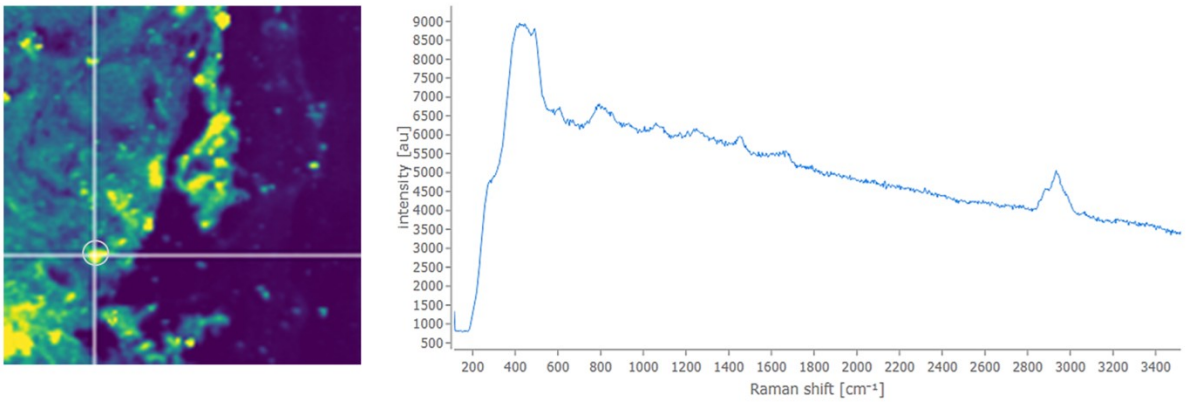


Figure 4.1: **Sample Raman spectrum from map 6c**: the effects of each pipeline step will be shown as an example on this specific Raman spectrum. False-color image and plotted spectrum are produced with Ramapp.

considered sufficiently reliable, so it was eventually decided to discard it. In addition, the full Raman spectrum has a peculiar shape such that the quick and sharp increase in intensity at a low wavenumbers, resulting from a combination of contributions, given by the signal belonging to the main peak produced by the quartz and autofluorescence baseline, tends to worsen the performance of the baseline correction step, thus giving a further reason in support of the removal of the wavenumbers up to 650 cm^{-1}. Instead, for values of Raman shift greater than 3050 cm^{-1}, only contributions from water molecules are usually expected to be seen; however, in this specific kind of samples, no particular signal from water can be collected and as a consequence, considering the absence of information, also this spectral region is being discarded from the analyses.

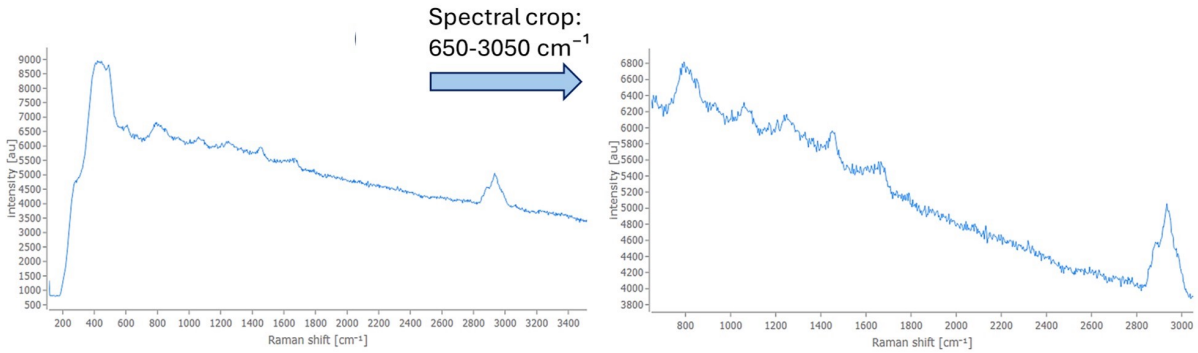


Figure 4.2: **Step 3 of the pipeline: spectral crop**.

The following step is the identification of cosmic rays signals, performed by means of a Z-score algorithm imposing a threshold equal to 7, a value which resulted to provide the

best trade-off between correct spikes identification and false positives occurrences. All the pixels that are marked for containing a spike produced by cosmic rays are corrected by replacing their spectrum with the median spectrum of its neighboring pixels.

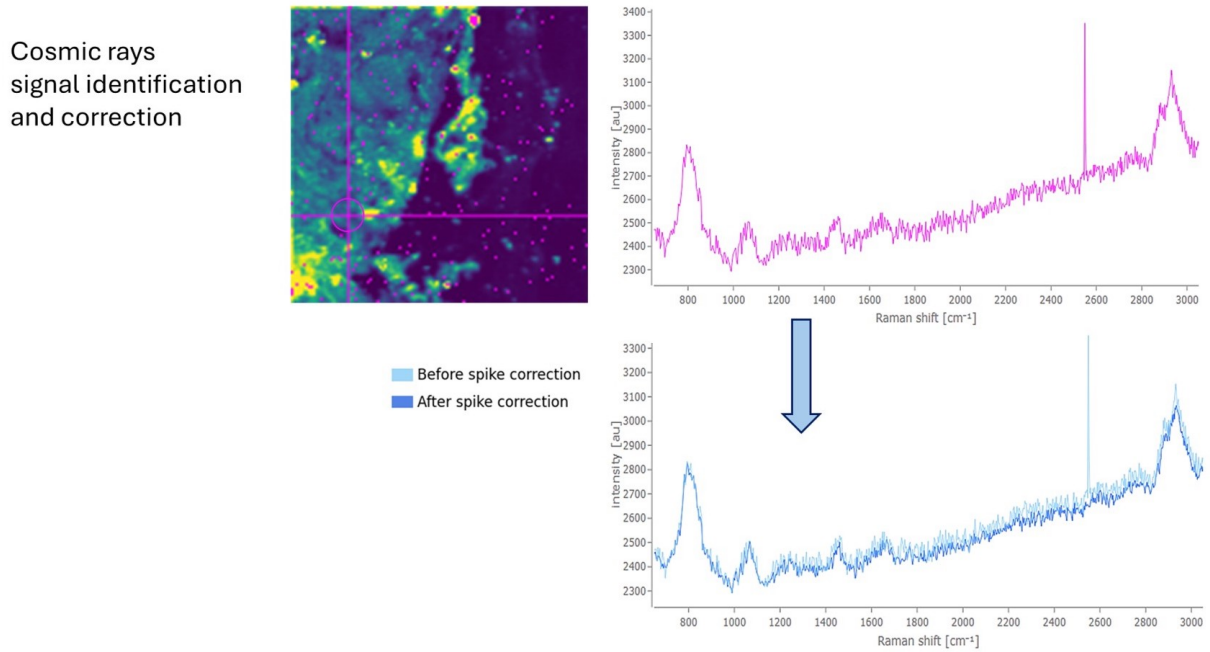


Figure 4.3: **Step 4 of the pipeline: cosmic rays signal identification and correction**.

The baseline correction is performed using the algorithm of Adaptive Smoothness Penalized Least Squares method (asPLS) [40], with the smoothing parameter $\lambda = 10^6$.

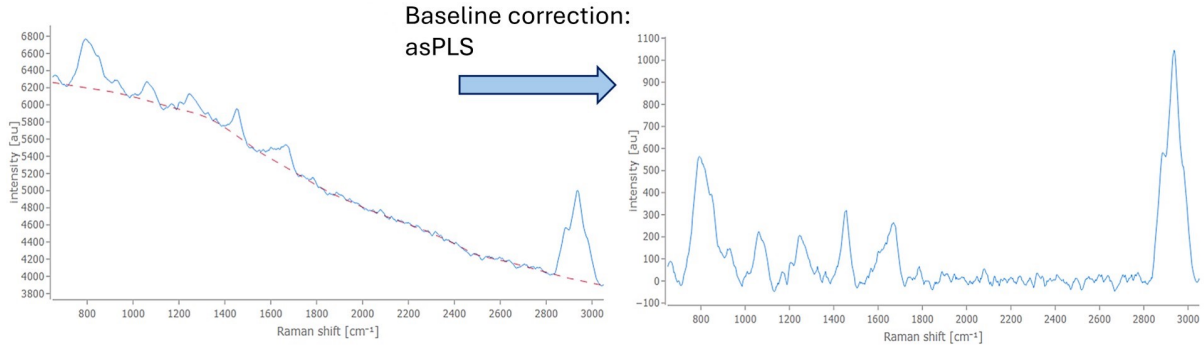


Figure 4.4: **Step 6 of the pipeline: baseline correction**. Adaptive Smothness Penalized Least Squares method with $\lambda = 10^6$

At this point of the preprocessing, one can conclude with a final rescaling of the spectrum with a normalization step, in order to remove the dependence on intensities by the acquisition parameters. One aspect that is taken in particular consideration is that, depending on factors such as the local thickness of the sample and concentration of Raman-active components, the ratio between the highest peak from quartz retained in the spectral region of interest and the peak in the CH region may change between values greater or smaller than 1; in other words, when the signal produced by biochemical components is not particularly high, the maximum intensity within the spectral region of 650 - 3050 cm^{-1} is reached by the quartz peak at around 795 cm^{-1}. As a consequence, if one performed a MinMax normalization, turning the intensities values into I' such that:

$$I' = \frac{I - \min(I)}{\max(I) - \min(I)} \tag{4.1}$$

the wavenumber corresponding to the maximum intensity value may change from pixel to pixel; we would like, instead, to have always the CH region peak to be identified as highest intensity one, as it would be expected in any Raman spectrum of biological samples, without interference from other contaminants contributions. Therefore, it is preferred to employ the L2 vector normalization:

$$I' = \frac{I}{\|I\|_2} = \frac{I}{\sqrt{\sum_{i=1}^{n} I_i^2}} \tag{4.2}$$

for which the impact of a different quartz contribution is still important but less disrupting for the following analyses.

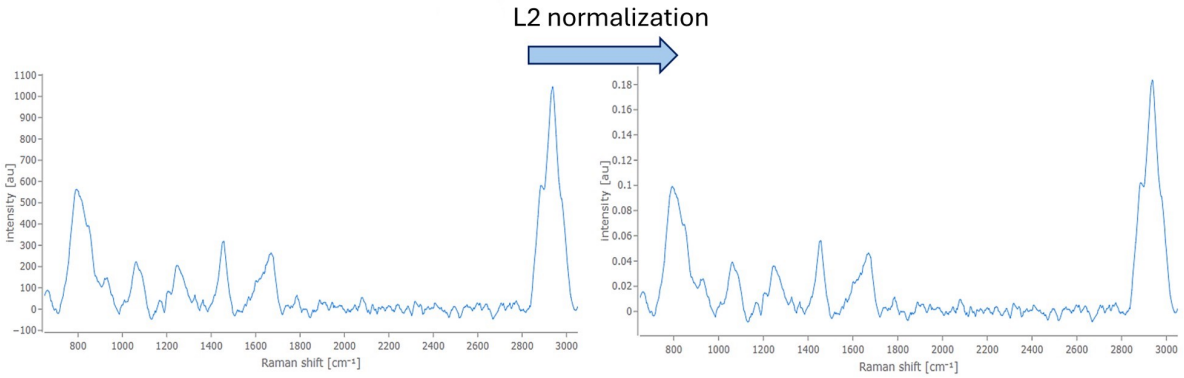


Figure 4.5: **Step 7 of the pipeline: normalization with L2 norm**.

Another option for the preprocessing pipeline has been explored, that is the introduction

of a smoothing step right after the baseline correction and before performing the intensity normalization. This choice is being supported by the chance of improving the signal to noise ratio, which can likely be beneficial for the analysis, especially in the case of lower quality signals. Indeed, the acquisition parameters have been tuned aiming at providing good quality spectra; however, the specific study that is being carried on requires the acquisition of hundreds of maps, each of them comprising 14400 Raman spectra. This is an aspect that cannot be ignored in the choice of the acquisition parameter, and forces the adoption of a pixel integration time that is compatible with this purpose, keeping the time required for data collection at a reasonably low value. Due to this limitation on the pixel integration time, and together with the significant variability on the measurements outcomes that stems from the intrinsic heterogeneity of the samples, the noise contributions become sometimes significant. On the other hand, attention must be paid to ensure that the smoothing algorithm does not alter the original spectrum covering existing peaks or reshaping it suggesting the presence of Raman signals which do not really exist, but are originated from a misinterpretation of noise terms.

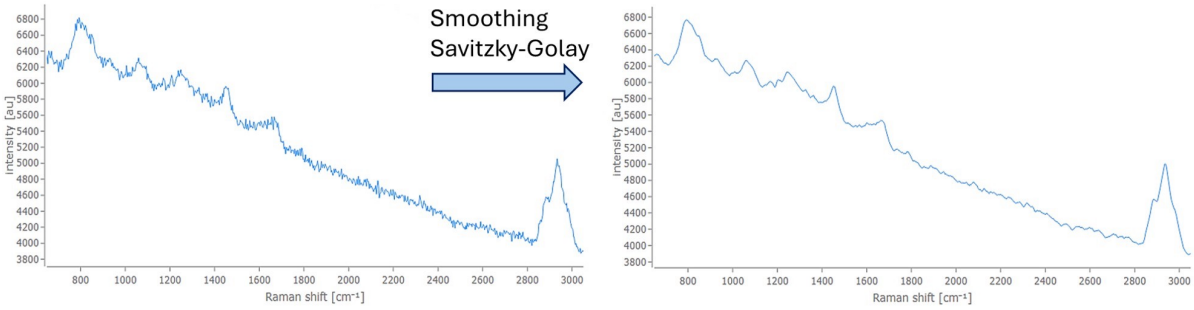


Figure 4.6: **Step 5 of the pipeline: spectral smoothing**: Savitzky-Golay filtering with window length 11 and polynomial order 2.

For the establishment of the preprocessing pipeline, two different smoothing algorithms have been tested to reduce noise: singular value decomposition (SVD) based on spatial ratio selection criterion, and Savitzky-Golay filtering [41].

Overall, a set of 3 pipelines has been tested, having in common almost all preprocessing steps but the fifth one:

1. X-axis interpolation to a set of 1024 values, equally spaced, from 200 to 3500 cm^{-1};
2. Map crop removing first row and column of pixels (from a 121x121 to a 120x120 map);

3. Spectral crop between 650 and 3050 cm^{-1};
4. Cosmic rays spikes identification and correction with Z-score method, threshold parameter set at 7;
5. Smoothing, which can be either one of the following:
   - SVD, spatial ratio selection criterion (pipeline 1);
   - Savitzky-Golay filter, with window length 11 and polynomial order 2 (pipeline 2);
   - 3: No smoothing algorithm, that is, spectra are directly normalized after baseline correction (step 5 is skipped) (pipeline 3);
6. Baseline correction with asPLS method, λ=10^6;
7. L2 vector normalization.

A summary of the pipelines outputs and comparison can be found in figure 4.7.

To evaluate them, the average spectrum relative to the biological sample is extracted by performing a k-means clustering restricted to the CH region (2800-3050 cm^{-1}), identifying 4 clusters and selecting as background quartz the one featuring the lowest peak centered at 2930 cm^{-1}. The average spectrum of foreground pixels of each map is plotted, and the associated signal to noise ratio is computed by considering the ratio between the area under the peak at 1660 cm^{-1} and the root mean square value in the wavenumber range of 2050-2150 cm^{-1}. For the latter, the choice of a small spectral range for the computation of the noise term was made due to the presence of some oscillations in the silent region around the expected 0 value; this feature can be mostly attributed to artifacts related to an etaloning effect produced at the camera. Those oscillations are of low intensity and the possibility to remove them by means of a more aggressive baseline correction had to be excluded, since also other Raman peaks would have been affected. Therefore, having verified that this shape is not originated from biochemical components, contaminants or suboptimal measurement conditions that could be readily fixed, and being its presence relevant only in the silent region, the final choice was to just account for it in the following analyses. Attempts are being made to reduce this spurious behaviour by means of background signal correction (see paragraph on EMSC) and proper adjustments are consequently considered for the analyses to come (for example, analyzing the CH region starting from 2820 cm^{-1}).

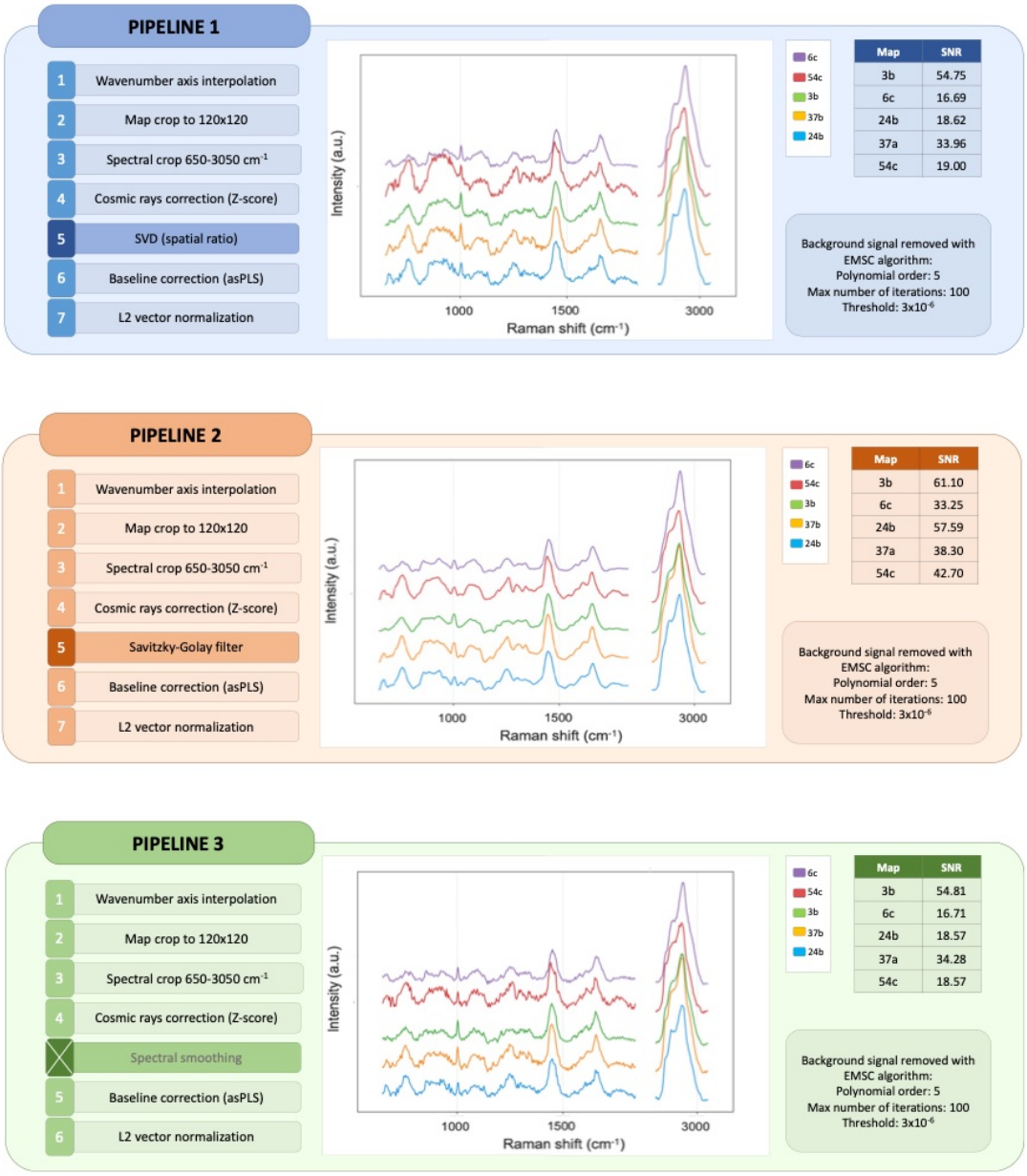


Figure 4.7: **Comparison between three different preprocessing pipelines**. For each of them, from left to right: the ordered steps of the preprocessing, the average foreground spectrum of every map after preprocessing, the SNR for each average spectrum after preprocessing. Average spectra are plotted after the application of a background signal removal algorithm (EMSC) which will be illustrated in detail in the following dedicated paragraph. The step is posterior to the full preprocessing pipeline and to the computation of SNR values reported in the tables.

For the comparison of the proposed pipelines, we start considering the average spectrum of foreground pixels of each map. We notice as anticipated an improvement in the SNR when applying Savitzky-Golay smoothing filter with respect to the case of not employing any smoothing algorithm. Considering for example map 24b, the average spectrum obtained with pipeline 3 without any map smoothing amounts to 18.57, while for pipeline 2 its value rises up to 57.58. What instead is distant from the expected behaviour is the very little improvement provided by the SVD algorithm, which is in addition not consistent among all the maps average spectra: only maps 24b and 54c show an improvement in the SNR values, while for 3b, 6c and 37a the value decreases. Improvements are in any case very limited: for map 24b, a SNR of 18.62 is obtained in contrast with 18.57 obtained without spectral smoothing. SNR values for every map average spectrum are reported in table 4.2.
Moving our attention instead to individual Raman spectra from each pixel, a reduction of the noise terms can indeed be identified. Table 4.1 summarizes the average SNR value computed over all foreground pixels of each map, to provide information on noise terms at single pixel level. It is possible to notice that SVD is more effective at reducing the contributions from noise with respect to Savitzky-Golay filter: if we name SNR_i the mean signal to noise ratio produced with pipeline i, for all maps of the test dataset we obtain that:

$$SNR_1 > SNR_2 > SNR_3 \tag{4.3}$$

However, comparing the average SNR value relative to individual pixels with respect to the mean spectrum of entire maps, the absence of a further improvement indicates a possible problem in the implementation of the SVD algorithm.
In particular, in case of uncorrelated noise, one would expect a reduction of the resulting standard deviation relative to noise σ_{avg} by a factor $\sqrt{N}$ with respect to the one of single spectra σ_i, with N number of samples included in the average:

$$\sigma_{avg} = \sigma_i * \frac{1}{\sqrt{N}} \tag{4.4}$$

This condition holds if and only if all the existing noise terms are uncorrelated with each other. In our case, this is not verified even in the case when no smoothing algorithm is applied; from this, we can deduce that some noise elements are not exactly entirely random but may be originating from specific sources. In any case, however, even if of lower magnitude than the one indicated in 4.4, an improvement is expected to be seen when passing from single spectra SNR values to the entire map average. This is eventually verified with the exception of the pipeline adopting SVD algorithm for noise removal.

This result, therefore, is not suggesting that no improvement is obtained by the use of SVD, but rather that the shape of the spectra is affected in such a way that also some noise features are preserved identically among adjacent pixels. As a consequence, when considering the mean spectrum, noise terms are not averaged out and the SNR ratio remains almost unaltered.

	24b	**37a**	**3b**	**54c**	**6c**
Pipeline 1	19.64	35.49	55.86	21.26	17.07
Pipeline 2	13.25	16.22	16.28	9.99	9.71
Pipeline 3	4.22	8.42	9.32	3.99	3.72

Table 4.1: **Average SNR of single pixels in each map**.

	24b	**37a**	**3b**	**54c**	**6c**
Pipeline 1	18.62	33.96	54.75	18.99	16.69
Pipeline 2	57.58	38.30	61.10	42.70	33.25
Pipeline 3	18.57	34.28	54.81	18.57	16.71

Table 4.2: **SNR of the mean spectrum of foreground pixels for each map**.

In conclusion, the final choice was to select pipeline 2, including the spectral smoothing step by means of Savitzky-Golay filter.
Pipeline 2 has been considered a better option because it produced, as expected, less noisy individual spectra, with an average SNR two or three times higher with respect to a pipeline not including any denoising step. At the same time, a good compromise in the selection of the algorithm parameters has been found, without any significant loss of information or deformation of the Raman features that could reduce or alter too much the information contained in the spectra.
Pipeline 1, involving the use of SVD, performs better in increasing the SNR. However, as it can be noticed when passing to maps average spectra, also some noise features are still retained and end up being shared in a similar way between the individual spectra included in each map. This effect has been considered as an introduction of bias due to improper application of the algorithm, which could be more detrimental than accepting a lower SNR value, thus leading to the selection of pipeline 2 over pipeline 1.

One final remark is required regarding the results provided by the pipeline using SVD.

Indeed, SVD is a very frequently adopted algorithm for noise components removal; specifically for this work, the preprocessing is performed with the same codes used by Ramapp application. In particular, the option of automatic selection of components that are to be retained or not has been adopted; the results suggest however that the choice is not optimal, producing the inclusion of components that are instead noise-related. Further investigation and refinements are planned for a more accurate selection criterion in order to evaluate SVD algorithm performances more accurately.

4.1.2. Custom made preprocessing tools

In the previous paragraph, some preprocessing pipelines based on Ramapp source code have been presented and compared using different metrics. However, some specific issues and needs for the collected dataset and case study are not fully addressed or tackled by this tool; to make the preprocessing procedure more complete and allow for a better preparation and filtering of the data to be analyzed, additional custom written Python codes have been produced, in particular implementing:

- the Extensive Multiplicative Signal Correction (EMSC) algorithm to remove quartz background signal;
- a global k-means cluster analysis, used both for post-preprocessing analysis and for substrate identification basing the classification on approximated estimate of the peak height in the CH region (see paragraph 4.2 for details);
- an automatic identification of pixels containing artifacts produced during measurements;
- a simple interactive interface to refine the classification of mapped tissues into healthy, tumoral or other areas as indicated in histopathologists' annotations.

EMSC for background signal subtraction

As seen in the previous paragraphs, one issue that is present in this case study is related to the strong signal produced from the quartz substrate on which the samples are provided. Removing the Raman contribution from quartz is not trivial; supposing that a separation between pixels referring to the background and pixels containing the biological specimen is achieved with a certain degree of accuracy, different possibilities are open. Ramapp codes allow the subtraction of the background spectrum with different approaches:

- Subtraction of the mean spectrum of the pixels classified as substrate;

- Subtraction of the median spectrum of the pixels classified as substrate;
- Subtraction of the mean spectrum of the pixels classified as substrate, comprising only those spectra which fall within the central 95% of values.

All these methods share the fact that the same spectrum is subtracted from all the pixels of the map, which is often enough to produce good quality results. In our specific case, the complexity and heterogeneity of the samples are extremely high, both from a morphological and biochemical point of view (different local thickness, cell density and biomolecular composition). Mainly for this reason, the shape of single spectra can vary significantly from pixel to pixel; as a consequence, the performance of the optical substrate removal algorithm is compromised, leading sometimes to an incomplete removal of the quartz signal or, on the opposite side, to the subtraction of an excessive intensity value with respect to the required one, with the production of negative intensity peaks.

An attempt to improve the background correction introducing a pixel-wise rescaling of the subtracted background spectrum has been made; in particular, the choice fell to the implementation of the Extensive Multiplicative Signal Correction (EMSC) algorithm, described in detail in [42]. Such algorithm requires a sample spectrum of the "contaminant" that is expected to be seen and which is aimed to be eliminated from the collected spectra, which is incorporated into a standard modified polynomial fitting method. The Raman spectrum, in particular, is modeled as composed by a set of peaks of Lorentzian shape, a broader lineshape accounting for fluorescence and the interferent spectrum. For each spectrum to be background-subtracted, a set of iterations is carried out to obtain the best estimation of the weight relative to the contaminant component, which is accordingly rescaled and subtracted from the measured spectrum. In our case, the reference spectrum that is provided to the algorithm is being extracted map by map as the average of substrate pixels, classified as such through Ramapp code which allows to perform the substrate identification, which performs a k-means clustering on a spectral region of choice, and selects the cluster with the lowest average value as optical substrate. For this specific purpose, the ROI selected is the wavenumber range of the CH region (2800 - 3050 cm^{-1}), and the number of clusters is chosen as the maximum available number (50) in order to reduce as much as possible the inclusion of foreground pixels in the background set. Even though the spectrum obtained does not exactly match the true average background signal, we are ensured that no features from biologically relevant signals are being included in the quartz reference spectrum to be rescaled and subtracted (so we are making sure not to remove contributions deriving from the sample). The algorithm, together with the required background identification step, is performed on already preprocessed data by means of pipeline 2 previously described. The other parameters to be defined

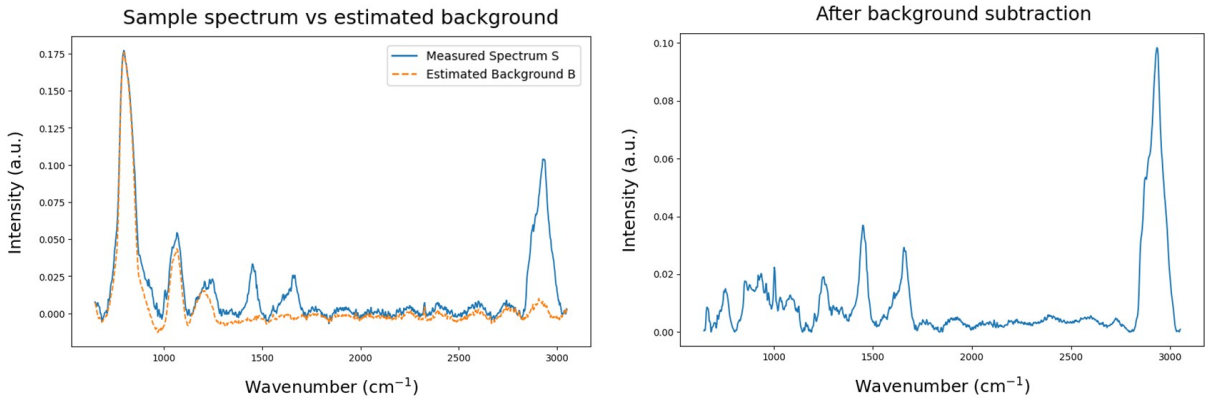


Figure 4.8: **Example of EMSC for background signal subtraction**.

consist in the maximum number of iteration Nmax and the polynomial order p for the estimation of the baseline background signal. For what regards the former, an increase in the maximum number of iteration corresponds to a generally better performance but also to a larger amount of time required to complete each process. The final value selected was 100. Considering the latter, instead, since the baseline has already been subtracted, no additional polynomial contributions should be present; the order chosen for the polynomial contribution was therefore tested both with value 0 and 5 (as presented in [42]). The results obtained in both cases are quite similar; a slightly more precise result is achieved by making use of polynomial order p=5, which was in the end chosen even if the time costs required are much higher than with polynomial order 0: assuming Nmax=100, a 120x120 map requires a total time of about 25 minutes for the background subtraction with EMSC. The resulting time for the entire dataset substrate removal amounts to 8300 minutes; indeed, the time required is high, and for sure improvements are possible, such as limiting the background subtraction just to foreground pixels. Still, the overall time cost is considered acceptable since it is a one-off operation.
An additional baseline correction is added after background subtraction (asPLS method, smoothing parameter $\lambda = 10^6$) as a final offset correction step.

Automatic identification of artifacts

An issue that occurred though the entire process of collection of Raman maps is the inclusion, inside the mapped areas, of artifacts-related signals. These are mostly associated with local heating and consequent burning of the sample; when this phenomenon is present, the Raman spectrum appears with a peculiar shape like the one shown in figure 4.9.

Such spectra completely lose information on the original biochemical composition of the

sample prior to the burning, and it is reasonable to aim at excluding them from the analysis.

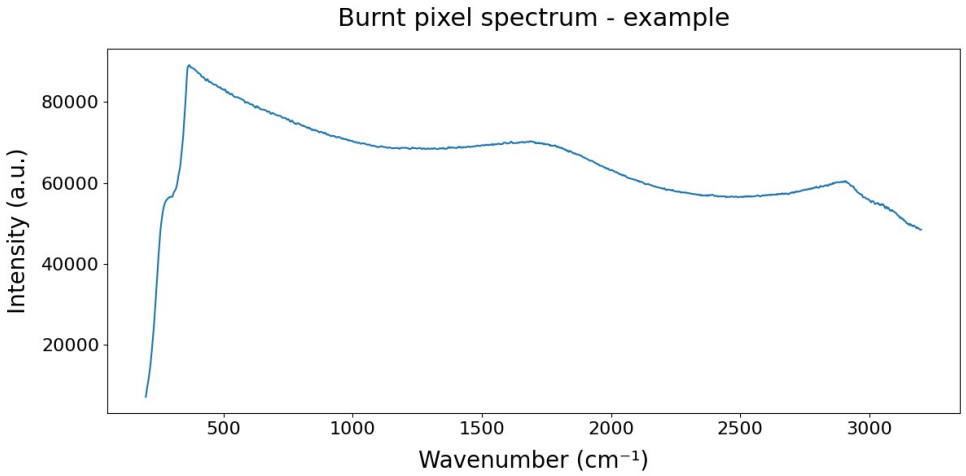


Figure 4.9: **Example of Raman spectrum in a burnt spot**.

Attention has been paid on many aspects to limit the occurrence of these episodes as much as possible; for example, in the selection of the region of interest to be covered by a map, samples regions including abnormally high cell density, or accumulations of tissue material due to scratching or folding of the slices are carefully avoided. In addition to this, the laser power and acquisition times are retained at a reasonably low value, choosing the best balance to ensure a satisfying quality of the collected spectra. However, in spite of some improvements thanks to thoughtful selection of mapped areas, this phenomenon could not be completely avoided; indeed, also local tissue composition plays a significant role, and the presence of biochemical components particularly prone to absorb light at the wavelength of irradiation cannot be easily inferred from simple observation of the sample under the microscope. The sample alteration caused by burning is most of the times localized in few pixels inside the maps, which may have an overall good quality in terms of collected signal; excluding entire maps for a feature affecting less than 0.01% of its pixels might be a too drastic approach, considering also that repeating the measurements whenever the phenomenon occurs would dramatically increase the time employed for measurements, alongside with the limitation of available ROIs on the sample for the specific region (healthy or tumoral) to be mapped. The use of the manual masking tool illustrated in the following paragraph has been considered as an option, but again this approach would be suitable for areas that extend to more than 4-5 pixels as instead is often the case; the production of ad hoc false color images, which would allow to easily identify the pixels affected by burning, is not only a non trivial task, but would actually mean to be somehow already able to identify the affected pixels with some sort of criterion,

making the introduction of a manual selection step redundant or even detrimental for the inherited risk of introducing operator-dependent errors. Finally, the presence of a manual step like this would be in contrast with the spirit of a data management and preprocessing that is as automatic as possible, also with the limitation of becoming unfeasible when the magnitude of the dataset scales up with the progress of the project.

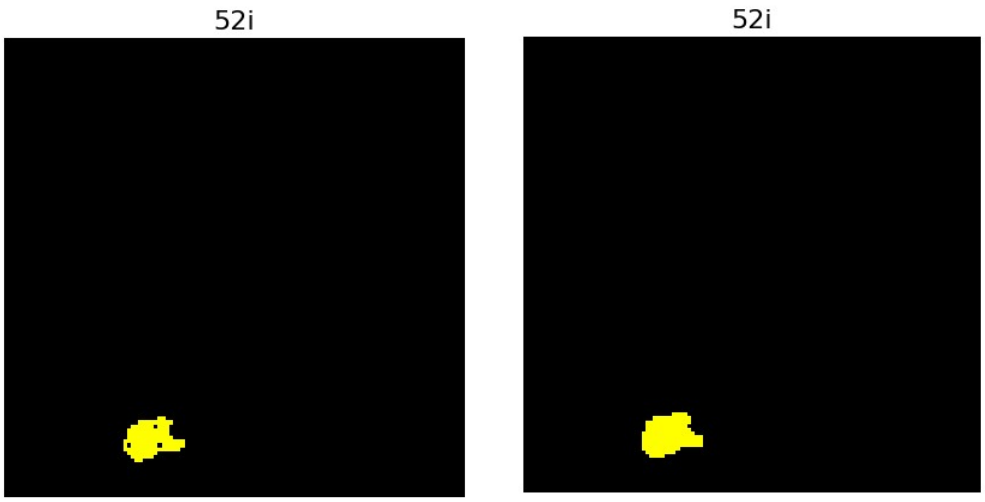
An attempt to automatically tackle the problem is here presented; in particular, two different features of the burnt spots Raman spectrum were exploited for their identification:

1. The higher signal intensity: burnt pixels feature a very high intensity Raman spectrum with respect to unaffected pixels, with maximum values in artifacts-associated pixels being up to 2 orders of magnitude larger;
2. The characteristic shape, in particular within the wavenumber region 200-1200 cm^{-1}, which, starting from the local maximum value that appears in between 350-500 cm^{-1}, recalls the typical curve of an exponential decay.

Considering these two characteristics, the algorithm is applied to non preprocessed data, firstly filtering all the Raman spectra of every map retaining only those which show an average value between 200-2000 cm^{-1} that is greater than 25000 counts. This value has been empirically chosen according to the typical values and orders of magnitude of the Raman signal intensity of the collected maps (under the assumption that the acquisition parameters are kept constant throughout all measurements). This first criterion alone, despite being rough and aspecific with respect to the actual shape of the signal collected, is already able to successfully spot all the burnt pixels present in the maps. The main limitation is that also pixels characterized by a strong autofluorescence baseline are included and removed; this is somehow still desirable when the fluorescence signal is such to cover entirely the Raman features of interest, since the effect would be a cleaning of the dataset from low informative spectra, but an additional control has been reckoned in any case as required to avoid the removal of acceptable quality spectra. For this reason, the choice went towards a more refined selection with a two-step method, based on the previously described intensity thresholding followed by an exponential fitting procedure. In particular, considering only those spectra which passed the first filtering step according to their average intensity, the maximum value in the wavenumber region 350-500 cm^{-1} is identified; the full spectrum is then cropped from the position of the maximum up to 1200 cm^{-1}, normalized with respect to the maximum, and then fitted with an exponential function. Only pixels achieving an R^2 value greater than 0.98 are retained. The normalization procedure is performed to allow a proper initial guess for the exponential function parameters, while the constraint on the R^2 value is quite loose due to the superimposed

background noise that can affect the outcome of the fitting. As a final step, pixels which are adjacent to at least 2 other pixels marked as burnt are then reclassified as burnt, to exclude those close to burnt areas which may still be affected by some unusual spectral features, which however do not strictly follow the typical one shown before in 4.9.

According to whether each pixel passed this two-step selection, binary masks are created for each map and exported as *numpy* arrays to be then imported and used in other codes dedicated to statistical analyses.
To design the algorithm and verify its proper functioning, the average spectrum of the removed pixels per every map is being automatically computed and plotted to allow a direct visual inspection from the user, together with previews of the masks generated and metrics summarizing the percentage of pixels in every map that have been marked as 'burnt'.



(a) Mask for artifacts (b) Final mask after propagation

Figure 4.10: **Example of artifacts masks for map 52i (patient SC23, day 2, Cisplatin 3,33 μM, mixed areas)** by applying the two-step selection method (a) and after reclassifying pixels adjacent to burnt ones (b).

Interactive interface for manual annotation

One last yet utterly important aspect that was being tackled is linked to the correct classification of pixels of every map into the respective tissue type classes indicated in histopathologists' annotations. In this work, two main classes are under major focus, namely, tumor tissue and healthy tissue. Occasionally, other categories have been included, specifically: necrosis and skeletal muscle (mostly for patient SC23), glioma, stroma and artifacts. The proper assignment of every map and every pixel to the corresponding

tissue type is fundamental for the further analyses that are to be performed. As explained chapter 3, the regions of interest to be covered by Raman maps are selected by a visual comparison between the brightfield images of the tissue slice to be measured and the H&E image provided with annotations referring to a contiguous tissue slice. Raman hyperspectral maps are collected in a specific, single-type tissue region and the whole map is classified accordingly as healthy or tumoral (other classes, except for necrosis in patient SC23, are avoided). A frequent issue encountered in this procedure is that it is not always possible to identify regions inside the measured tissue slice with a unique classification and with an extension comparable or greater than the field of view covered by one map (i.e. 600x600 μm^2). In some other cases, such "entirely healthy" or "entirely tumor" areas can be found; however, being able to select with more flexibility the region of interest, including also multiple classes inside a single map, would allow to investigate areas with peculiar morphological features, select an area providing more intense Raman signal or avoid regions of the tissue slices that are expected to produce artifacts.
This procedure is a common standard for many studies, yet more precise assignment can be reached if information are provided on the exact same tissue slice measured with Raman. This piece of information is expected to be provided in the future but is not yet available for the study presented in this work.

The proposed tool is therefore to be intended as a temporary solution, suitable for when a limited number of maps needs to be masked, and only for those cases in which the separation between different areas is well-defined and possible to distinguish inside the Raman map. The problem of correlating the spectral information contained in Raman maps and the information provided by histopathologists by establishing a pixel-by-pixel correspondence is being planned as one of the next steps within the project work, in concomitance with the accessibility to staining and annotations on the same exact measured tissue slices.

To help refining the classification of some maps' pixels, a simple, interactive user interface has been developed with the use of *tkinter* Python library. The code allows to manually draw masks on a map of choice (the "main image"), that can be selected and displayed on the interface. At the same time, it is possible to also select and display a second image file besides the main image, to be used as a reference to trace separation lines (therefore called "reference image"). The tool allows to select whether the sides of the main image are automatically considered as margins of masks, to ease the procedure when the drawn line starts from one side and ends on another point on the perimeter of the main image. Once a mask is being defined, the user is asked a name describing the area

selected (e.g. tumoral, healthy, necrosis...) and a numerical label (integer number of any value, 0 included). Multiple masks can be created on the same main image, also using labels already selected for previous masks, which may be needed when a mixture of tissue types are present inside the same map with a more complex distribution than a two-sides separation. When all the desired masks have been drawn, the user can terminate the process and one last time prompts appear asking for a name and a numerical id to be assigned to all the pixels that have not been masked. According to the selections made, a 2D *numpy* array file is saved with the following characteristics: the dimensions of the array correspond to the dimensions in pixels of the uploaded main image. Each element of the 2D matrix corresponds to the integer number assigned to the mask, while the correspondence between numerical id and tissue type indicated in the process is stored in a .csv file. These files can be accessed later by the data analyses codes to update the pixel by pixel assignment to the corresponding histopathologists' area classification.

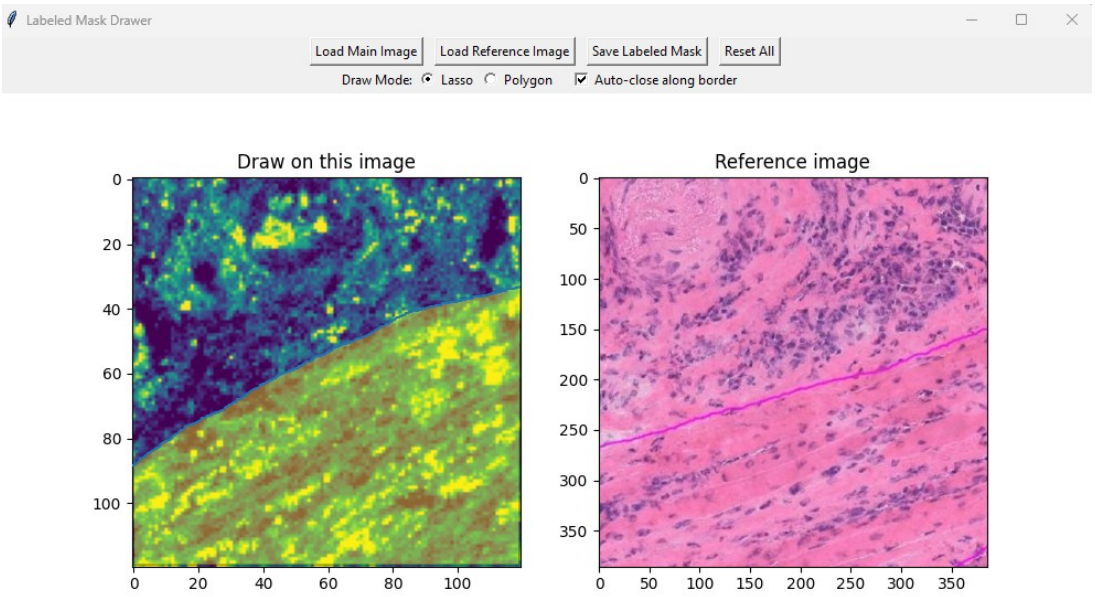


Figure 4.11: **User interface for manual annotations on Raman maps**. The tissue below the traced line, highlighted in yellow, is classified as skeletal muscle tissue, the one above more generically as healthy. The reference image is the H&E stained contiguous slice, approximately in the same area covered by the map, with the purple line drawn by histopatologists separating skeletal muscle tissue (below the line) and the rest of the covered field of view.

In order to have consistency between the size of masks produced and Raman maps, the "main image" selected is always chosen as a false color image of the Raman map after

interpolation and map crop (first two steps of the pipelines illustrated in fig. 4.7) so that a fixed mask dimension of 120x120 is ensured, simplifying a lot the following procedure without the need of any further manipulation on the masks' array dimensions. False color Raman maps are produced optimizing the visualization according to the average intensity in correspondence of the CH region, to better highlight the contrast between foreground and background and have clearer resemblance with possible reference images of any kind; any other kind of manipulation or choice is possible according to the specific visualization need. No particular constraints are applied with regards to the choice of the reference image, which can be of any kind (brightfield image, stained H&E contiguous slice with or without annotation, etc.), of any size and including a region of interest that can be arbitrarily larger or smaller with respect to the main image Raman map; uploading this image is just meant to be an aid in the annotation process and can either be used or not according to the user's preferences. An example of this tool's interface is provided in figure 4.11.

All the tools described in this paragraph have been integrated in the workflow as schematically illustrated in figure 4.12, including the main steps regarding the acquisition and the preprocessing procedure. Data analysis codes are finally provided with:

- masks of artifacts-related pixels, to be excluded from the analysis. To preserve the peculiar shape of damaged pixels, maps used for the creation of these masks have not been fully preprocessed, but just interpolated and cropped to a size of 120x120, to ensure that their dimensions in pixel number matched the one of fully preprocessed ones and facilitate the integration of this information in the data analysis steps;
- the hyperspectral Raman maps preprocessed according to pipeline 2 and background-subtracted through the EMSC algorithm;
- masks indicating pixel-wise tissue type classification, according to the manual annotation performed. This input is not provided for every single Raman map, but only for those that presented an heterogeneous composition and could not be attributed to a single category.

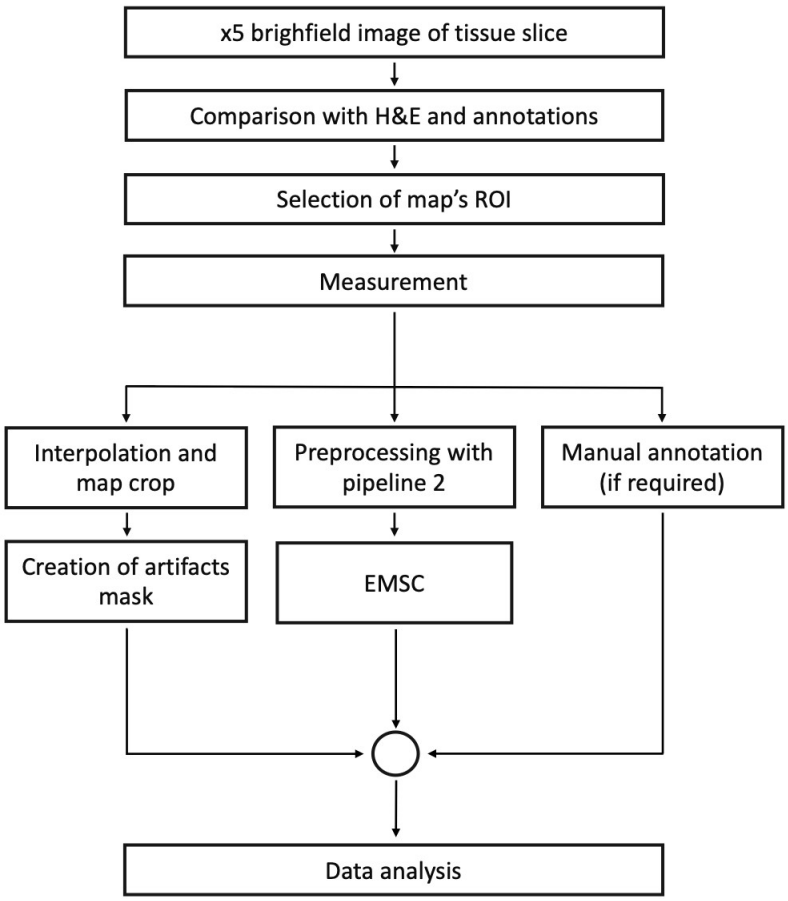


Figure 4.12: **Schematic summary of acquisiton and pre-analysis steps**.

4.2. Raman maps analysis

The dataset of hyperspectral Raman maps at the present day amounts to a total of 332 images of 120x120 spectra. For the analyses that will be presented in this work, the choice was to select a subset of the available data according to the following criteria:

- accuracy of maps area classification and correspondence to H&E annotated images;
- residual substrate quartz signal.

Regarding the first point, every Raman map needs to be either assigned to one specific category of tissue (e.g tumor, generic healthy, skeletal muscle, necrosis...). In case a Raman map covers a region including more than one tissue type, it is necessary to make sure that all the categories can be properly separated by means of the manual masking tool. This aspect is of great importance for the analyses that are being carried out, since the majority of them will be based on comparisons and classifications that either aim at discriminating tissue classes or require precise information for the selection of the set of maps to perform the analysis, which must be suitable according to the specific purpose of the investigation. For example, if an interested in treatment effects on tumor tissue is present, it is fundamental to dispose of accurately classified tumor tissue area, since spoiling the dataset with points belonging to a different category may hinder the results or lead to incorrect and biased conclusions.
Maps that do not meet this condition are excluded from the analyses presented in this thesis work. In particular, the impossibility to provide a classification with acceptable reliability is in most of the cases related to the differences between the H&E annotated images and the mapped tissue slice. Indeed, contiguous tissue slices are very similar to each other, but do not always present exactly the same structures and shape; moreover, though less frequently, also some deformations, scratching or small tears are produced during the manipulation of the tissue slices in the sample preparation. All these factors, together with the intrinsically low contrast of brightfield images, can sometimes prevent the establishment of a proper correspondence between measured area (every map is always located precisely inside the tissue slice) and annotated region.
In other cases, instead, localizing the acquired map in the H&E annotated image is possible but the features and distribution of the different tissue types are too complex to be reproduced manually with the manual annotation tool. Examples of these are shown in figure 4.13.

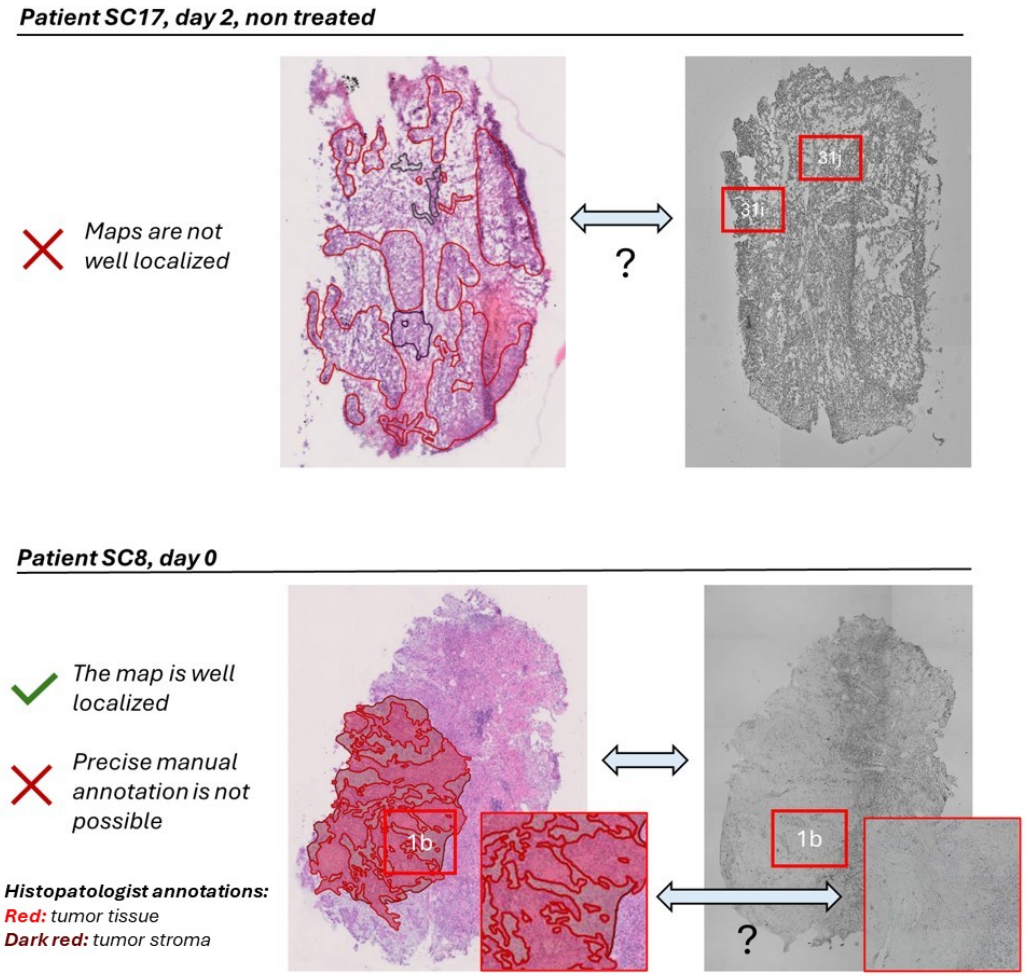


Figure 4.13: **Examples of maps temporarily excluded from the analyses.**

The second aspect regards the effectiveness of the background signal subtraction and the general quality of the spectral data. In particular, in few cases the removal of the quartz signal does not produce optimal results and spectra still show some residual substrate features at low wavenumbers (650-1100 cm^{-1}). The occurrence has been noticed to be linked also to overall lower spectral quality of the collected map. Since substrate contributions inside the spectra together with enhanced noise features can bias the analyses (especially clustering approaches and PCA), for the present work maps affected by this phenomenon are not entirely discarded but are included just in those analyses where only the CH region or a limited portion of the fingerprint region is investigated.

Table 4.3 summarizes the final dataset considered for this work, which includes a total of 52 maps. The accurate classification of maps collected on patient SC15 and SC17 was particularly challenging due to the difficulty of recognizing on measured slices the patterns present in the H&E images, essential to guide the manual separation of different tissue areas on the false color Raman maps. In addition, for these two patients only preliminary annotations were available at the time when analyses have been performed, with discrimination only of tumoral tissue with respect to non tumoral one. For this reason, it was preferred to keep a conservative approach by including only a reduced number of maps. The full dataset of acquired hyperspectral images is planned to be entirely included in future analyses as soon as more refined annotations will be available, possibly on the very same measured slices.

Patient	SC7	SC8	SC15	SC23	Total
Total number collected maps	13	14	70	39	136
Unsure classification	5	5	50	15	75
Low signal	0	0	6	2	8
Total available maps	8	9	14	22	53

Table 4.3: **Final dataset for analyses**.

4.2.1. Analysis of average foreground spectra

K-means clustering for background identification

The very first kind of analysis that is being performed on Raman hyperspectral maps is focused on the average spectrum associated to foreground pixels. In particular, the goal is to obtain some preliminary information on the local biochemical tissue composition, in particular in the light of the methodology and preprocessing pipeline adopted. To perform this kind of analysis, the first step is to separate between foreground and substrate pixels. Indeed, this operation has been included in the previous preprocessing steps with the aim of extracting a sample average spectrum of quartz, to be then used as input to perform the background signal removal with the EMSC algorithm. However, considering this scope, the process has been characterized by a conservative approach in terms of reducing as much as possible wrong classifications of foreground pixels in the background-associated cluster. Indeed, including biochemical Raman features inside the average quartz spectrum

that is going to subtracted is strongly undesired. This need has been considered of prior importance with respect to a precise separation of tissue and substrate areas and led to the choice of a high number of clusters (k=50); information related to single pixels assignment to the relative clusters is eventually discarded.
The context of this analysis is instead different, requiring an accurate classification both of foreground pixels and background pixels in their respective classes. For this purpose, a custom made code has been implemented with the functionality of performing k-means cluster analysis, taking as example the tools available on Ramapp. In the same way as in the web app, the k-means clustering can be performed with a free choice of the spectral region of interest, that can be either one specific wavenumber range, or the entire spectral window with the option of ignoring the silent region, reducing the risk of pixels being grouped by features that are only related to noise. The code allows also to perform a local normalization in the spectral ROI selected (L2 norm or MinMax).

The novelty introduced with this code is the possibility of performing the analysis and applying the algorithms "globally", meaning that all the spectral data inside the dataset of choice are processed together. Operations are therefore no more performed at a single map level, allowing for example to create clusters that are built and shared among all the maps, or to apply PCA or other multivariate analysis tool on a larger set of maps and finding components that express the variance of the data not only in one specific map. Finally, in addition to this, before performing the clustering algorithm or any other operation of interest, artifacts masks produced with the dedicated code described in 4.1.2 are integrated in the process in such a way that pixels classified as artifacts-related are directly excluded from the very beginning. In this way, any interference at every step of the data analysis, starting in first instance from foreground pixels identification.

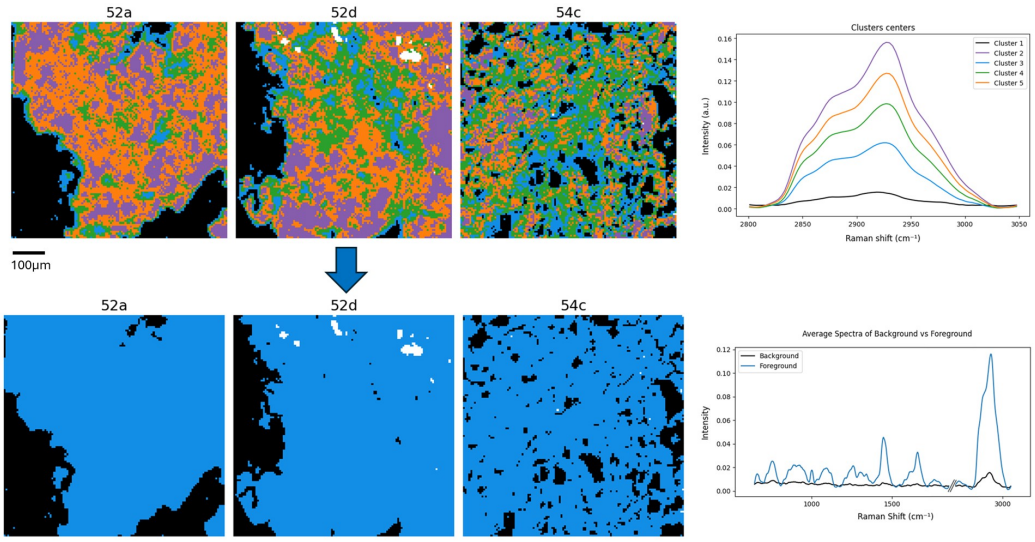


Figure 4.14: **Schematic representation of the foreground identification process**: maps are first classified into 5 clusters, then the cluster featuring the lowest peak is automatically labelled as background. After the process, two masks are retained: foreground (blue) and background (black). White pixels, excluded from k-means clustering, are directly imported as masks related to artifacts.

Considering specifically the application of k-means clustering for separating foreground pixels from background ones, the choice was to focus only in the CH region and working with 5 clusters, without any additional normalization. Indeed, the spectral window of 2800-3200 cm^{-1} is a good candidate for separating biological samples from non-biological ones, since C-H bonds widely present in biomolecules produce a characteristic peak with a much higher intensity with respect to fingerprint features.
What can be observed for example from figure 4.14 is that, with this choice, clusters differences are mostly related to intensity; this differentiation has been exploited by labeling as "background clusters" the one whose centroid shows the lowest CH peak, as it is expected in Raman spectra not produced from biomolecules. All the pixels that are instead associated to one of the remaining 4 clusters are retained as foreground pixels. The final output consists therefore in the pair of foreground and background mask for each map.
Figure 4.14 shows the process for 3 Raman maps of patient SC23: the k-means clustering separates pixels by intensity, areas characterized by a stronger biological signal (linked, for example, to a greater thickness of the tissue or density of cells) are assigned, in increasing order, to cluster 3, 4, 5 and 1. Cluster number 2, presenting the lowest peak in the CH region, has been associated to quartz substrate. After the identification of

the background, all other pixels are generically classified as foreground ones and the finer subdivision in each specific cluster is discarded.

Maps average spectrum and differences

After having identified the pixels of interest belonging to the sample, the average signal of each Raman map can be computed. As mentioned in the previous paragraph, not all maps can be associated to one single tissue type; in such cases, averaging all the pixels and classifying uniquely the average spectrum obtained could be an overly simplified approach, as features that may be specific and characteristic to each particular type of tissue would be blended together and obstaculate following comparative analyses.
Let's for example consider map 7b, measured on the patient SC7 non-treated slice at timepoint 2 and shown in 4.15. The region of interested covered by the map includes a certain heterogeneity in tissue composition, including tumor tissue, tumor stroma, blood vessel and skeletal muscle tissue. By making use of the manual annotation tool, the 2 regions related to the blood vessel and skeletal muscle were separated from the rest of the map, which has been approximated in first instance as tumor since it was not possible to identify a clear distinction of the two of them. By computing separately the average spectrum of each manually annotated region (excluding background pixels, while artifacts are not present in the map) one obtains the three spectra shown in fig. 4.15. Differences between them are instead plotted in 4.16.

Muscle and blood vessel tissue difference (first plot (a) in fig. 4.16) is not characterized by very sharp and defined peaks, but it is possible to identify generally greater protein-related contributions, such as in amide III (1220 - 1300 cm^{-1}) and amide I (1600-1660 cm^{-1}) bands and consistently with the positive peak in the CH region at 2930 cm^{-1} (see fig. 1.7 as reference). As we will see later, when performing kmeans cluster analysis, pixels of the blood vessel tissue are partially assigned to the cluster mostly characterizing muscle tissue, while the others to the ones characterizing stroma or epithelial tissue, thus suggesting some shared features with both tissue types and less marked differences.
On the other hand, moving to the following two plots b and c of fig. 4.16 we see that tumor tissue is characterized by stronger proline and collagen Raman signals (at 815 cm^{-1}, 856 cm^{-1}, 920 cm^{-1}, 1163 cm^{-1}) with respect to both blood vessel tissue and muscle tissue. Also proteins signal at 1275 cm^{-1} and amide II band at 1480 cm^{-1} are more intense, while the signal related to amide I band is instead reduced. Decreasing intensity is also registered at 1003 cm^{-1} and 1607 cm^{-1}, which can be associated to phenylalanine. Finally, particularly considering the comparison between tumor tissue and blood vessel tissue, contributions from DNA related signal are also more intense, for example

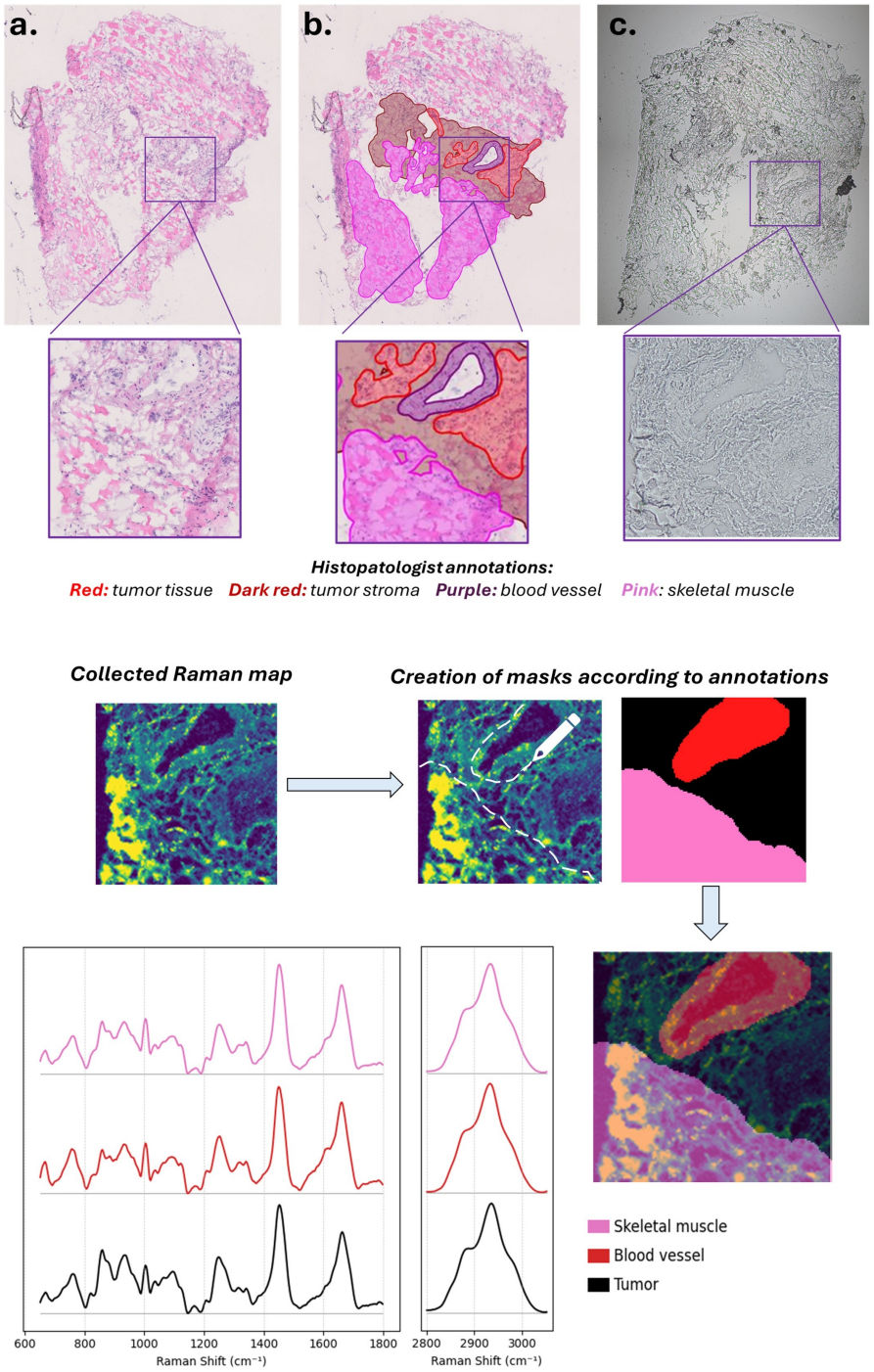


Figure 4.15: **Averages between different areas of map 7b**: 3 different areas can be identified on map 7b, belonging to patient SC7, day 3, non-treated. Here reported the position of the map on (a) H&E stained continuous slice, (b) annotated H&E image and (c) brightfield image of the measured slice. Average spectra for each area are plotted with an added offset and scaled to facilitate visualization.

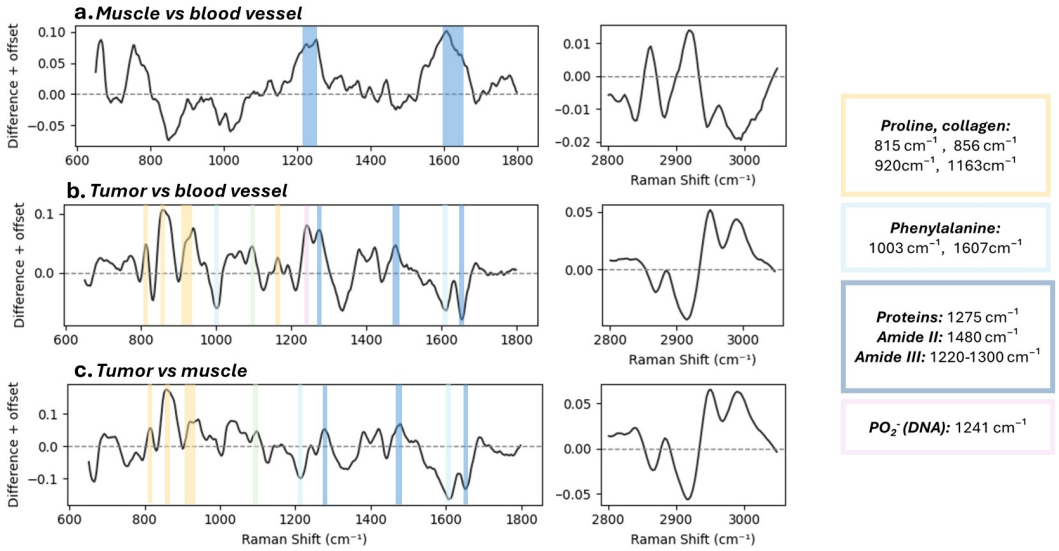


Figure 4.16: **Differences between average spectra of tumoral tissue, skeletal muscle and blood vessel tissue in map 7b**.

the asymmetric PO_2^- stretching at 1241 cm^{-1}. Such results can be interpreted considering that tumor cells growth and multiplication are usually upregulated with respect to healthy cells, making it reasonable to find cells in a more densely packed organization. in addition, tumor cells observed in H&E images present a larger nucleus and generally higher nucleus to cytoplasm ratio. These aspects may explain the reason why an average stronger contribution from DNA-related components is retrieved in cancer associated tissue with respect to blood vessels and skeletal muscle.

This can make us conclude that, as expected, it is more reasonable to treat separately the different tissue regions, whenever this is made possible by a sufficiently resemblance of morphological features of measured slice, annotated slice and acquired Raman map.

To separate foreground and substrate pixels and compute the average spectra, maps have been processed in batches including all maps belonging to the same patient; for instance, k-means clustering has been applied on the set of maps from patient SC7, foreground pixels have been identified, then separately for every tissue type the average Raman spectrum has been calculated. The same procedure is repeated for all other patients samples. Figure 4.17 provides an example of the averages of patient SC8 tumor tissue maps, with the exceptions of maps 1a and 1b which, as illustrated in figure 4.13, show a complex composition of tumor and tumor stroma that cannot be separated manually.
By looking at the spectra in fig. 4.17, some of the main peaks associated to biological signal can be identified. Indeed, the CH region features the usual peak whose shape results

from the convolution of contributions from lipids, proteins and nucleic acids. From one map to the other, some peaks may appear more evident, such as around 2885 cm^{-1}, when lipids provide major contributions, or the 2930 cm^{-1} main peak, when instead proteins are present in larger amount; signals from nucleic acids are not strongly visible at their main wavenumber contribution (2956 cm^{-1}), which is however typically expected as their presence is usually limited with respect to the other biomolecules. In the fingerprint region, instead, it is possible to identify peaks relative to amide I band (1640-1680 cm^{-1}), cytosine (1610 cm^{-1}), proteins (1453 cm^{-1}), lipids (1255 cm^{-1}, 1130 cm^{-1}), tryptophan (1208 cm^{-1}), collagen (937 cm^{-1}), proline and tyrosine (855 cm^{-1}) [43].

The silent region has not been plotted since it lacks of relevant information, and to improve visualization the CH region and fingerprint region are plotted separately, to allow a rescaling and a greater visibility of the signals within the range of 650-1800 cm^{-1}. Indeed, as widely reported from literature, the CH region peak intensity is dominant with respect to the fingerprint region signals.

From the stacked plots of fig. 4.17, it is already possible to see some subtle differences between peak intensities from one tissue slice to the other, such as different ratios between 1255 cm^{-1} and 1315-1343 cm^{-1} peaks, the greater and lower prominence of 1610 cm^{-1} peak, the different shape of the CH peak according to greater (4a, 4b) or lower (5a, 5b) contributions from lipids. This pushes for a deeper insight and comparison, therefore differences between spectra are computed and plotted. Unfourtunately, only one map covers a generic healthy tissue area that can be properly separated from the surrounding tissues of other kind; for this reason, the choice was to investigate differences between treated and non treated tumor areas rather than healthy and cancerous tissue maps.

Differences in time of treated and non-treated samples average spectra

One important question for our case study is whether providing cultured slices with PD-L1 protein is producing any kind of effect on cancer cells with the passing of time and especially if these changes can be detected to some extent from the Raman scattering data collected. Since, as already mentioned, maps 1a and 1b are characterized by a mixture of tumor tissue and tumor stroma, comparisons with timepoint zero are avoided considering the risk of introducing biases in the results due to differences in area composition. The focus is for this reason restricted to the differences between the average spectrum of treated samples at timepoint t4 and timepoint t2.
Results are being shown in figure 4.18; in particular, with the current dataset, there are

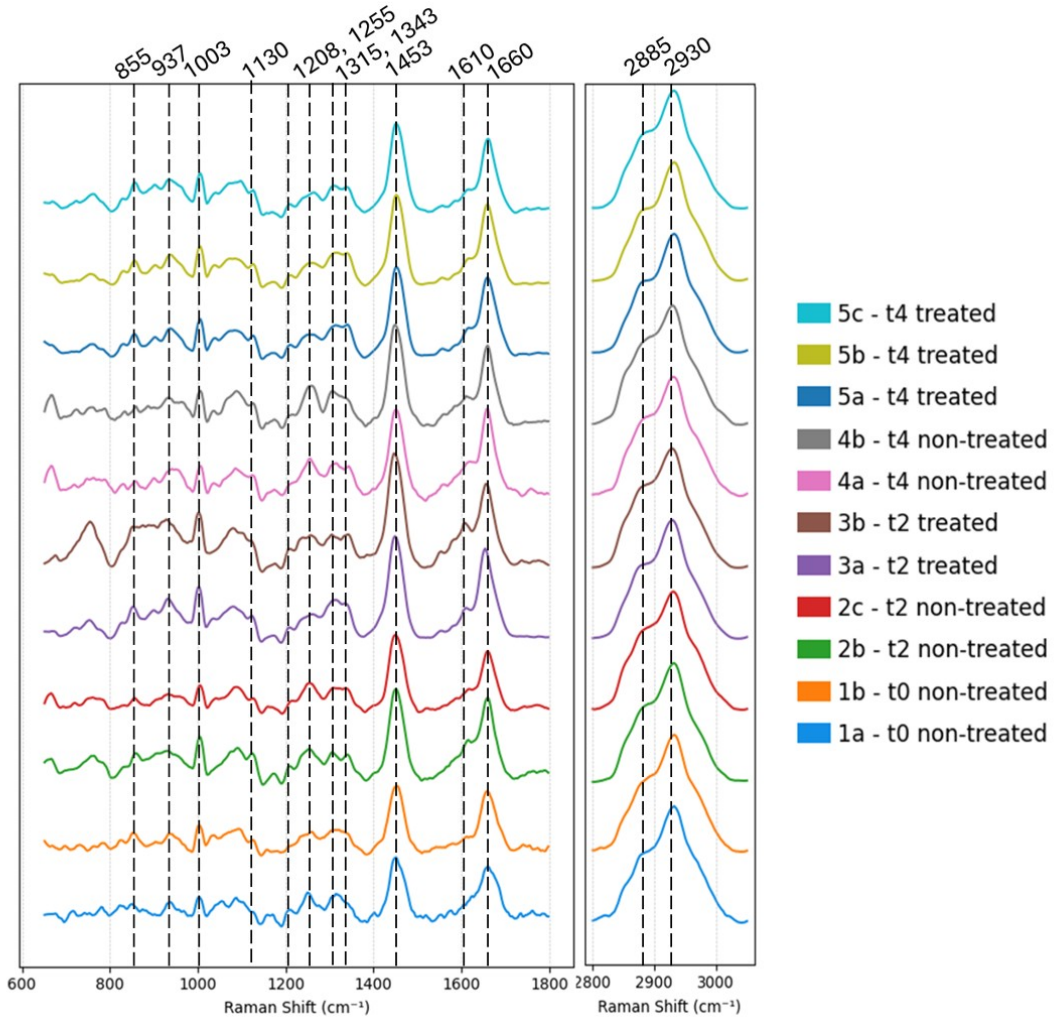


Figure 4.17: **Average spectrum of SC8 tumor tissue maps**.

3 maps available for timepoint 4 (5a, 5b and 5c) and 2 maps for timepoint 2 (3a, 3b); all the possible combinations for the subtraction are performed, to make sure that the conclusions are more general (at least for the patient SC8) and not specific to the choice of the maps to be subtracted.

From 4.18 we can see that some spectral features are common to all the pairs of maps, in particular:

- tumor tissues at timepoint 4 always show an increase in the CH region of the contribution of proteins (2930 cm^{-1}) and DNA signal (2956 cm^{-1}), while the wavenumbers associated to lipids follow an opposite trend of decrease;
- in the fingerprint region, signals of amide I band (1660 cm^{-1}) and amide II (1478 cm^{-1}) are stronger at timepoint 4; phenylalanine (1003 cm^{-1}) and lipids (1437

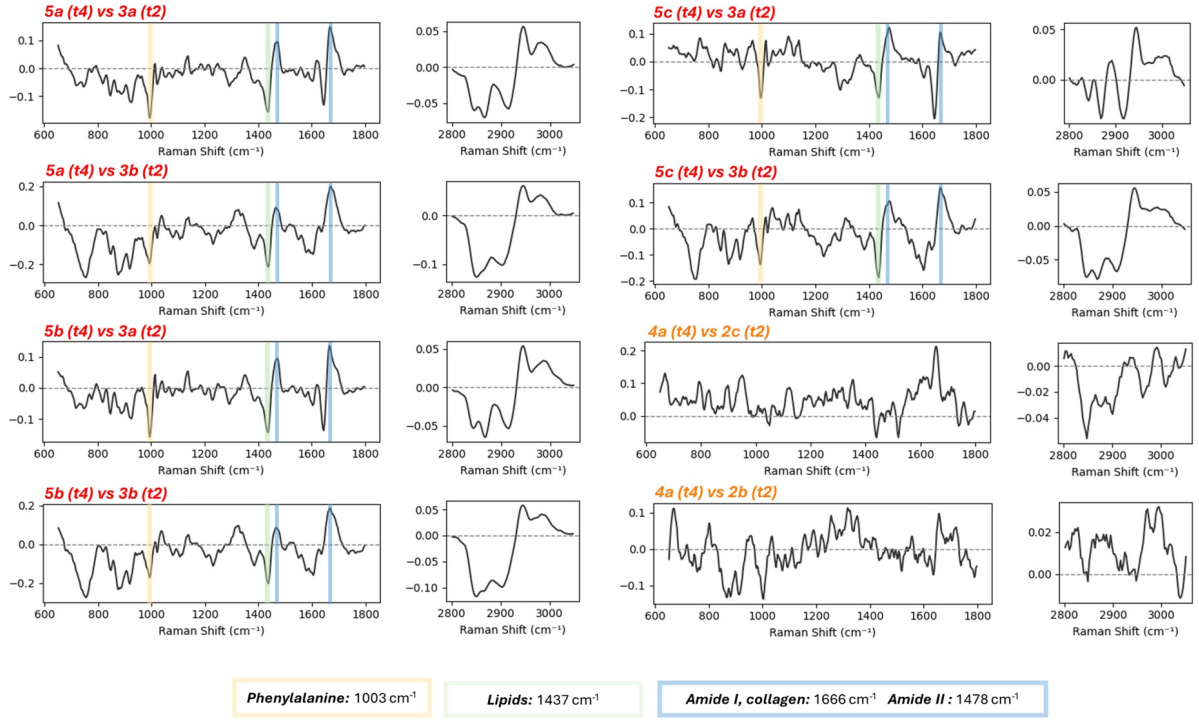


Figure 4.18: **Average spectrum differences of treated and non-treated tumor tissue maps in patient SC8**: maps 5a, 5b, 5c, 3a and 3b belong to treated tissue slices (red labels), while maps 4a, 2b and 2c have been collected on non-treated tumor tissue (orange labels).

cm^{-1}) on the contrary are lower.

These results, especially those related to increase in the DNA signal from the CH region, may suggest a general increase in cell density in tissues at day 4 with respect to day 2, which could be interpreted as an increased proliferation of tumor cells. This behaviour would be compatible with the working principle of the PD-L1 protein treatment that is being applied: as explained before, the treatment for patient SC8 consists in providing PD-L1 protein to the cultured tissue slice, in such a way that the antibody binds to PD1 receptors on immune cells, with the overall effect of reducing the immune system response. This would be coherent with the increase in the DNA content, which can result from an overall increase of cell density in the mapped tissue, stemming from a less restrained multiplication of cancer cells due to a depressed immune response.

The fact that this behaviour is associated specifically to the effect of the treatment is supported by the results from the average spectra difference for non-treated tumor maps, which have also been reported in 4.18. The average spectrum of map 4a (day 4, non-treated) is subtracted from map 2b and map 2c (both non-treated, at day 2) and such a trend of DNA content increase, with a shape as clear as for treated samples, is not

present. This preliminary result can therefore suggest that, specifically for patient SC8, samples are responding to the treatment in some way, showing a different evolution in time with respect to non-treated samples.

4.2.2. PCA analysis

Moving now to multivariate analysis, the application of Principal Component Analysis (PCA) is being explored. PCA is a widely spread tool for analysis of spectral data, and also in this specific study it has great potentiality for gaining deeper insights, allowing first of all to extract those spectral features that mostly explain the variance within the dataset; eventually, the retrieved components can be associated with biochemical interpretations, and may also highlight features that particularly contribute to separate classes related, for example, to tissue types, treatment condition or temporal evolution. However, some critical aspects have been identified in the application of PCA in this specific work:

- **Heterogeneity and complexity of sample composition.** PCA is employed in this work with the aim of separating classes which are used to label entire maps and not single pixel spectra. Each map, however, comprises many elements of great diversity that constitute biological tissues, and as a consequence a significant portion of the variance captured by PCA arises from point-localized structures. This means that the principal components maximize variance both between and within classes, thus limiting the ability to effectively separate the classes themselves. For the same reason, it cannot be excluded that truly discriminative spectral features are not present in every single spectrum of a map, but is characteristic only of some specific structures or areas inside the tissue.

- **Dataset dimensions.** The number of data points to which PCA is applied is not intrinsically a problem for the algorithm performances; however, this is an aspect to be taken into account in the visualization of the output provided by the tool. Indeed, 2D and 3D scatterplots are widely used and very powerful to provide an intuitive and immediate overview on data distribution (which is also one advantage provided by dimensionality reduction, i.e. the possibility of representing data at a lower dimensional level). Things however may change when a large dataset is used as it happens in this work: maps include 14400 Raman spectra each; if one would consider a PCA analysis performed on just 2 maps, and the case (very far from being representative) of having half of the pixels belonging to the tissue, scatterplots would include more than 10000 points. Such high numbers can easily hinder the clarity of the scatterplots, unless classes are very sharply separated along some principal

components, which however is considered quite unlikely to happen for tissue maps.

- **Single pixel spectrum quality.** The heterogeneity of the sampled area increases the complexity of PCA also because the algorithm works at a single spectrum level, meaning that it receives as input all spectra of pixels that are classified as foreground. It may happen that some pixels signal is characterized by lower signal to noise ratio; this usually happens to pixels belonging to thinner tissue regions, to those located at the boundary between sample and quartz, or even inside quartz regions close to the sample containing some biological material detached from the slice.

Considering the abovementioned aspects, PCA analysis has been applied for the moment on relatively limited number of maps.

In order to tackle the third point illustrated, i.e. the presence, even inside the best quality maps, of low signal-to-noise ratio Raman spectra, an additional clustering step is performed after substrate identification and before the application of PC decomposition. The aim of this step is to create clusters which include the lower quality pixels, allowing to discard them easily by removing entire clusters and clean the dataset from the most noisy elements.
In order to achieve this, 12 clusters are produced with foreground pixels data; this time, the full spectral window from 650 cm^{-1} to 3050 cm^{-1} is taken into account, and MinMax normalization is performed. Clusters related to noisy spectra are recognized either because they are characterized by a marked offset from the zero value of the entire centroid spectrum, or because the fingerprint region appears with an unusual shape, with possible residuals of the quartz peaks and again a positive offset (not necessarily present also in the CH region).
For the analyses presented in this work, the selection of the clusters to be discarded usually comprises 3-4 clusters out of 12, and corresponds in most of the cases to the 3-4 clusters including the lower number of points. Indeed, during the process of excluding a portion of data from the analysis, it is important to verify that the dataset is not excessively depleted, with the risk not only to miss important features but also of not being anymore representative of the original dataset. Before proceeding, the total number of pixels that are going to be removed is always verified and compared to the total amount of available foreground pixels.

PCA for treatment effect monitoring - Patient SC8

One first analysis that has been performed involves the same maps of patient SC8 whose average spectra have been compared in the previous paragraph.

Map name	Patient	Area	Timepoint	Treatment
4a	SC8	tumor	day 4	non-treated
4b	SC8	tumor	day 4	non-treated
5a	SC8	tumor	day 4	treated
5b	SC8	tumor	day 4	treated

Table 4.4: **Dataset for PCA analysis on SC8 maps**.

In particular, 2 maps of treated tumoral samples at timepoint 4 (5a and 5b) and 2 of non-treated samples at the same timepoint (4a and 4b) are investigated (table 4.4)to see if coherent results are obtained from mean spectrum differences and characterization based on the dimensionality reduction.
First, a k-means clustering in the wavenumber range 2800-3050 cm^{-1} with k=4 clusters for separating foreground and background pixels has been applied; a further clustering on foreground data is performed on the full spectrum for noisy signal removal, as described in the previous paragraph; the total of removed pixels amounted to about 10% of the foreground pixels. Finally, on the foreground pixels only of retained clusters, principal component analysis is performed:

- in CH region only, within the wavenumber range of 2820 - 3020 cm^{-1};
- in the two spectral from 1200 cm^{-1} to 1800 cm^{-1} and from 2800 cm^{-1} to 3050 cm^{-1}, ignoring the signals of the silent region from 1800 cm^{-1} to 2800 cm^{-1}.

In both cases, Raman spectra are MinMax normalized before applying PCA, and the number of principal components is set at 8.

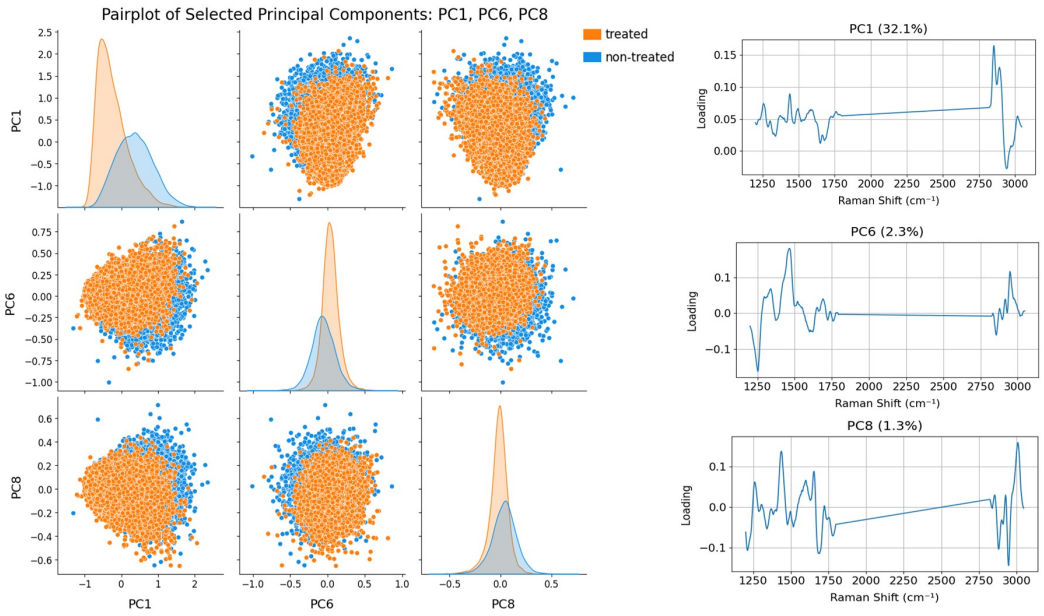


Figure 4.19: **Principal components 1, 6 and 8**: scatterplot of principal components number 1, 6, 8 and relative plots with explained variance indicated in parenthesis.

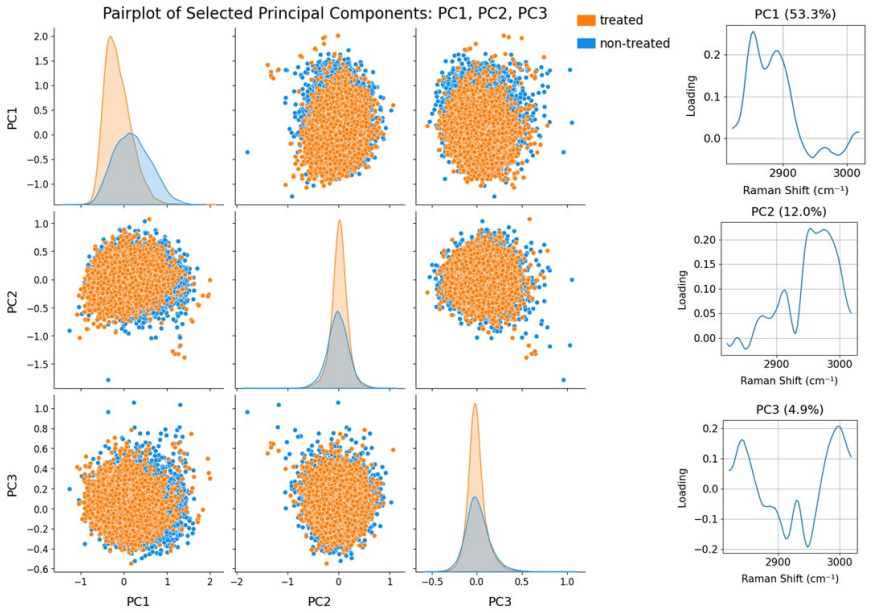


Figure 4.20: **Principal components 1, 2 and 3 (CH region)**: scatterplot of principal components number 1, 2, 3 and relative plots with explained variance indicated in parenthesis.

Specifically, for the **CH region** the first 4 principal components can be retained as informative in terms of biochemical interpretation, with particular interest for PC1, related to lipid content, and PC2, linked to nucleic acids signal with a minor lipid component contribution; PC5, PC6, PC7 and PC8 are instead to be considered as noise terms and have been excluded. The first three components scatterplots and the corresponding loadings are shown in 4.20. Separation between the two classes is slightly suggested by PC1, while scores of PC2 and PC3 seem to be identically distributed between the non-treated and treated samples. The behaviour relative to PC1 is in any case in agreement with what have been found previously: compared to non-treated tissue slices, tumor areas subject to treatment feature a lower contribution from lipids and an higher signal from DNA (overall prevalence of negative scores of PC1 for treated samples).

Regarding the **1200-3050 cm^{-1} spectral region** analysis, one might be questioning why the selected wavenumber region comprises just a portion of the fingerprint one. Indeed, PCA analysis on the full wavenumber range 650-3050 cm^{-1} has been performed; however, some PC components were observed to be still affected by some baseline residuals, and separation between points were provided only by the first principal component; results relative to this specific trials were considered of limited significance and have therefor not being included in this treatment. Removing an additional cluster of pixels was considered not appropriate as it would have significantly impacted the number of dataset points, and it was instead preferred to give up on the spectral information in the wavenumber region mostly affected by the phenomenon. The output from the restricted analysis eventually produced some components which better separated the two classes; therefore, despite being far from an optimal approach, the compromise of neglecting a portion of the fingerprint region was finally accepted.

Separation between the two classes is slightly suggested by the scores of PC1 in the CH region (fig. 4.20) , and by PC1, PC6 and PC8 in the full Raman spectrum (fig. 4.19) . As anticipated before, scatterplot are not very effective in conveying a possible separation of the data due to the very high amount of points. For this reason, also boxplots of PC scores of the principal components of interest are presented in 4.21 and 4.22, aiding at highlighting for example the differences in mean values of PC scores.

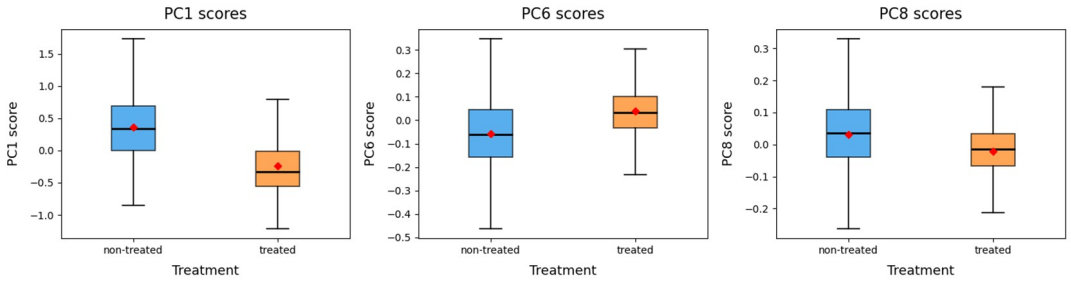


Figure 4.21: **Boxplots of PC1, PC6 and PC8 scores on full spectrum PCA analysis**. Shapiro-Wilk test provided low p-values ($p<0.001$), indicating that data do not follow a normal distribution. Mann-Whitney-U test was performed and all pairs resulted to have significant differences ($p<0.001$).

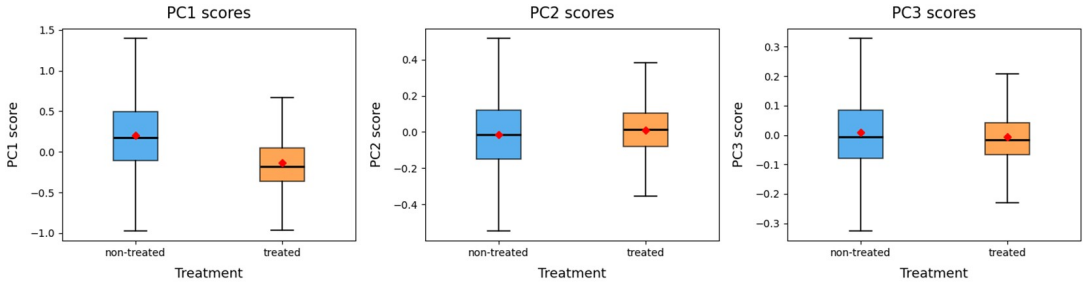


Figure 4.22: **Boxplots of PC1, PC2 and PC3 scores of CH region PCA analysis**. Shapiro-Wilk test provided low p-values ($p<0.001$), indicating that data do not follow a normal distribution. Mann-Whitney-U test was performed and all pairs resulted to have significant differences ($p<0.001$).

By looking at the results, the analysis is consistent with what has been found in the previous paragraph, suggesting that tumor regions in treated samples feature a lower lipid content than non-treated ones, but a greater presence of nucleic acids signal:

- from figure 4.22 we see in the CH region smaller PC1 scores, with a negative average, but higher PC2 scores and positive average;
- in figure 4.21, referring to full spectrum analysis, negative PC1 scores and positive PC6 scores are seen in treated samples, while an opposite behaviour is shown for non-treated samples.

PCA for treatment effects monitoring - PD-L1 treatment

A similar analysis based on PCA has then be performed on a larger dataset comprising maps from all the patients that have been treated with PD-L1 protein, namely patients SC7, SC8, SC15, and SC17. The idea is to follow a similar approach as the the one

illustrated for patient SC8, that aims at comparing treated and non-treated samples at the same timepoint and eventually identify possible features from PC components that might be descriptive of the patient response to the treatment. The dataset is the one reported in table 4.5 and comprises, for each patient, one treated map and one non-treated map covering at least a portion of tumor tissue, for two different timepoint to allow also the investigation of eventual evolutions in time, for a total of 4 maps per patient. SC17 is the only exception, for which only one timepoint has been used due to unsure classifications or lower quality signal of late timepoint maps.

Map name	Patient	Area	Timepoint	Treatment
2c	SC8	tumor	day 2	non-treated
3a	SC8	tumor	day 2	treated
4a	SC8	mixed	day 4	non-treated
5a	SC8	tumor	day 4	treated
7a	SC7	mixed	day 3	non-treated
8a	SC7	mixed	day 3	treated
9a	SC7	mixed	day 4	non-treated
10a	SC7	tumor	day 4	treated
12f	SC15	tumor	day 1	non-treated
13e	SC15	tumor	day 1	treated
14a	SC15	tumor	day 2	non-treated
15a	SC15	tumor	day 2	treated
32a	SC17	tumor	day 2	non-treated
33a	SC17	tumor	day 2	treated

Table 4.5: **Dataset for PCA for PD-L1 treatment response on 4 patients**. When the map covers a region with an heterogeneous tissue composition, the area classification is indicated as "mixed"; for the analysis, only signal from tumor-associated pixels has been considered.

The analysis is performed on pixels classified as tumor tissue and is focused in the CH region. Since the dataset is composed by a considerably high number of points, scatterplots produced by combination of principal components were produced but did not allow an easy and visually effective interpretation of the results; for this reason, the scatterplots not reported in this document. Instead, boxplots of the principal components scores are

shown in 4.24, with attention towards principal components 1 and 2 (fig. 4.23), that can be associated to lipids signal and DNA respectively.

What is possible to see is that there is not a trend in PC values that is shared exactly between all patients. In particular, we can make considerations by observing:

Treated and non-treated samples, at same timepoint:

- PC1 scores are lower in treated tissues with respect to non-treated samples in patients SC8 and SC15, while they are higher for patients SC7 and SC17;
- PC2 scores are lower in treated tissues with respect to non-treated ones in SC7 and SC8, while they are higher in SC15 and SC17.

Temporal evolution in time:

- For patients SC8 and SC15, PC1 scores decrease in time, with a more marked difference for treated samples. Patient SC7 shows instead an opposite behaviour but of more limited entity;
- PC2 scores increase in time for patient SC15 both for treated and non-treated samples, with a greater change in treated tissue. Also patient SC8 shows a slight increase in PC2 scores for treated samples, while as in PC1 the values remain very similar for non -treated samples. In patient SC7, a decreasing trend is seen in both treatment conditions.

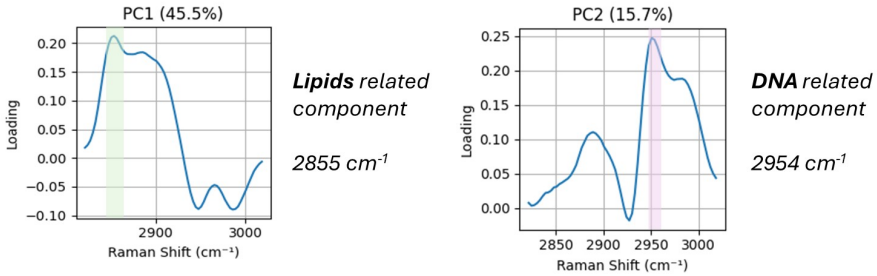


Figure 4.23: **Principal components 1 and 2 in CH region**.

Indeed, a decrease in lipid content paired with an increase in the DNA signal can be associated to tumor proliferation: the malignant cells multiplication can lead to a greater cell density and therefor to a greater collection of Raman signals from cell nuclei with respect to lower density areas. Tumor development is also often related to an increase in the presence of more protein components and an altered lipids metabolism, from which a reduction of lipids signal could be justified. Since the treatment provided is expected to act on the sample with a reduction of the immune system response, the trend seen in

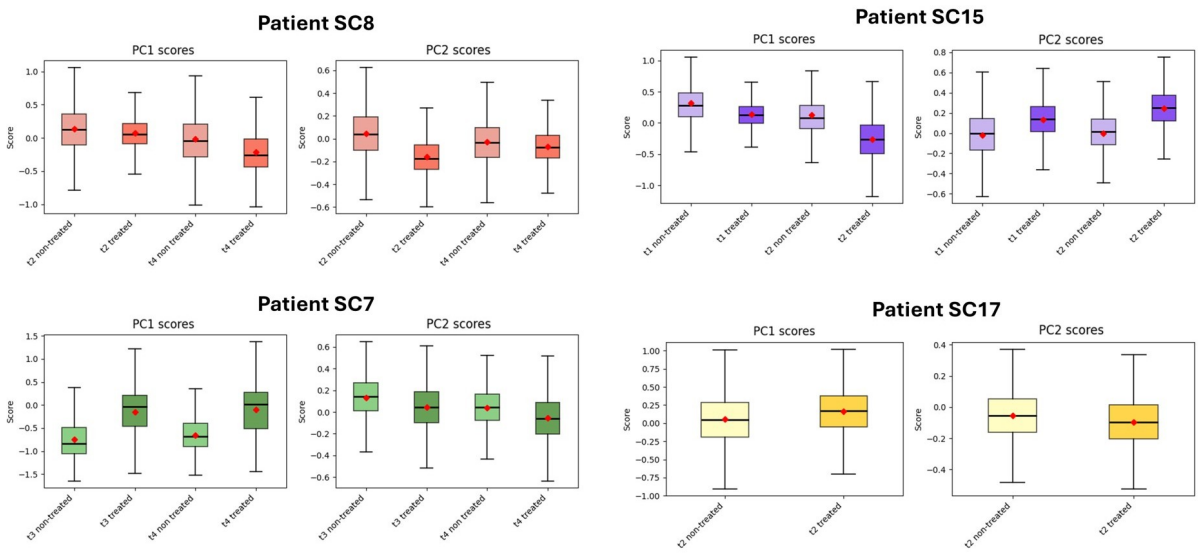


Figure 4.24: **Boxplots of PC1 and PC2 scores**: PC1 (left) and PC2 (right) scores for each patient, timepoint and treatment condition (lighter color shade for non-treated samples, darker for treated ones). Kruskal-Wallis test indicated significant differences across the groups (p-value < 0.001). Post-hoc pairwise comparisons using Dunn's test provided p$<$0.001 for all pairs with the only exception of PC2 scores of treated samples at day 3 and non-treated samples at day 4 of patient SC7.

of increasing scores for PC2, accounting for a stronger DNA contribution, and decreasing scores for PC1, indicating a reduction of lipids-related signal, in treated sample is reasonable.

Patient SC15 clearly presents according changes in time. Even if the same behaviour in time is present also non-treated ones, the magnitude of the changes is greater for treated areas, suggesting that what is observed is not just a characteristic of the tissue sample evolution during cultivation.

Also patient SC8 follows this behaviour; in addition, PC2 appears to be inverting its trend also when comparing non-treated to treated samples.

SC7 and SC17, however, seem to behave in a different way, with inverted trends with respect to those of SC8 and SC15. Indeed, tumor tissue response to treatment is a complex process and analyses carried out in a limited spectral region may not be enough to describe exhaustively the phenomenon, potentially missing some other key components that instead may be driving factors in the processes occurring. On the other hand, there are many other variables that have not been illustrated and included in this work which can significantly influence the response to treatment. In this framework, it is important to also acknowledge the possibility of observing different behaviours from patient to patient. In this context, it is interesting to notice that both patient SC7 and SC17 have been

diagnosed the same cancer type, namely the oropharynx non-keratinizing squamous cell carcinoma, and sharing in addition HPV-16 positivity. This is a an important element not to be overlooked, since SCCs associated to HPV positivity are known to exhibit distinct differences from HPV-negative HNSCC in gene expression, and mutational and immune profiles [17], with studies showing the peculiarity of non-keratinizing oropharyngeal carcinoma [44]. Patient SC8 and SC15 resulted instead HPV negative; patient SC8 has been diagnosed with a keratinizing tongue margin carcinoma, while patient SC15 tumor was a floor of mouth carcinoma. The correlation of PCs scores with HPV positivity would certainly require dedicated in-depth analyses and is therefore not been investigated in this work; however, it is reasonable to expect different responses of affected patients due to the peculiarity of this head and neck cancer subtype.

Finally, it is important to mention that the illustrated results are not to be intended as general or conclusive in the assessment of the overall treatment response; it is reasonable to believe that, due to the inherent complexity of the process, a description by linear dependencies alone may not be exhaustive and resolutive. Availability of more timepoints would indeed offer better insights on the real temporal evolution of samples and strengthen the robustness of the findings of this analysis, involving at the moment just two timepoints. The inclusion of a larger cohort of patients is indeed mandatory for generalizing the discussion and providing a foundation for validating the observed changes as descriptive of the temporal evolution of samples according to treatment.

4.2.3. K-means clustering for tissue type identification

Single patient tumor and tumor stroma separation

At the beginning of this chapter, the average spectra of different tissue regions inside one Raman map have been computed and compared with the calculation of their corresponding differences (fig. 4.15). For this operations to be performed, a manual annotation on the measured Raman image was required to label each pixel of the map to the correct area class. However, many examples in literature show the capability to follow an opposite approach, for which starting from Raman data and making use of algorithms of various kind, for example those based on clustering approaches, it is possible to identify regions in the field of view which are characterized by differences in their biochemical composition, without prior knowledge or labeling of the data. If the spectral features separating the identified groups result to be characteristic of the specific tissue type, this would allow the production of false color images based on cluster assignment which resemble the annotations on H&E stained tissue slices.

In this work, this possibility has been tested with a set of maps belonging to patient SC7. In particular, maps 6a, 6b, 6c collected on the untreated slice at timepoint 0 in different regions have been selected (fig. 4.25). By looking at the contiguous slice annotations, most of the covered areas can be classified as tumoral tissue; however, for maps 6a and 6c (fig. 4.25, on the left) also tumor stroma areas are included, which causes their exclusion from other analyses such as the one previously described on average map spectra, requiring a precise distinction among different tissue classes. The manual annotation in this case is not easy to implement due to a very scarse distribution in map 6a, while in the case of map 6c the main limitation is linked to the impossibility to recognize with clarity the different structures.

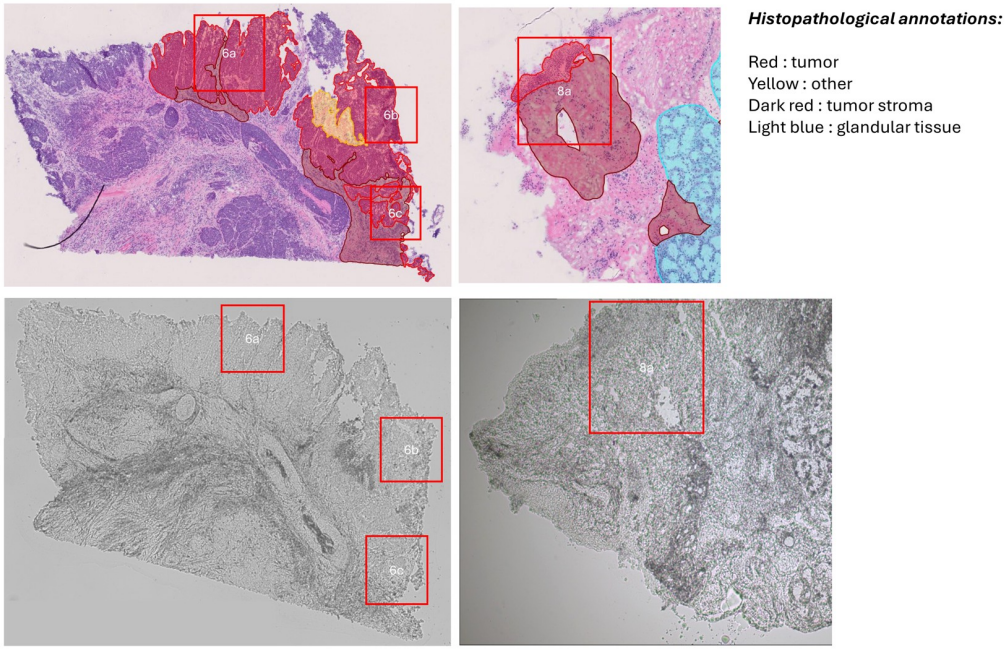


Figure 4.25: **Positions of maps 6a, 6b, 6c and 8a**: above, their localization on H&E annotated contiguous slice, below on the brightifield image of the measured tissue slice.

Map name	Patient	Area	Timepoint	Treatment
6a	SC7	mixed	day 0	non-treated
6b	SC7	tumor	day 0	non-treated
6c	SC7	mixed	day 0	non-treated
8a	SC7	mixed	day 3	treated

Table 4.6: **Dataset for tumor and tumor stroma K-means clustering.**

Maps 6a, 6b and 6c have been processed together with map 8a, located on the tissue slice of day 2, treated, for which a very clear separation between tumor stroma and tumor tissue can be seen from the comparison with annotations (fig. 4.25, right side).

K-means cluster analysis is applied on the 4 maps in the usual region of interest between 2800-3050 cm^{-1} for substrate identification. A subsequent k-means clustering is applied once more in the CH region, limitedly to the foreground pixels and performing a MinMax normalization.

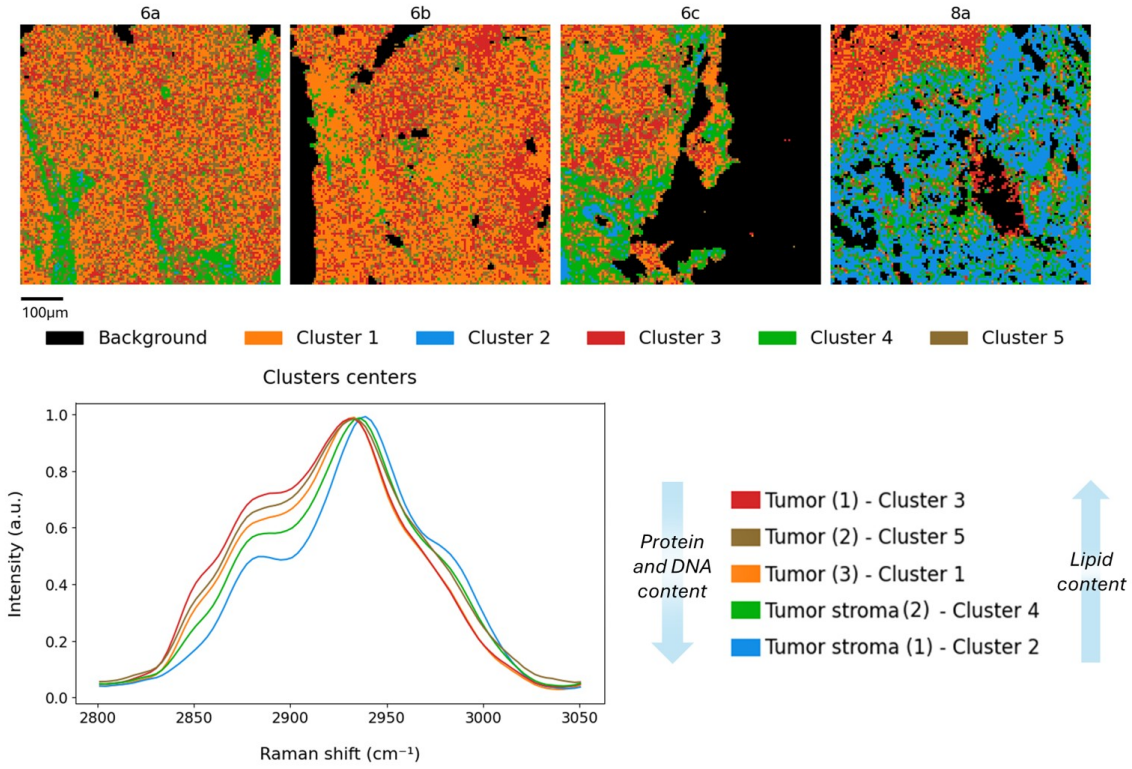


Figure 4.26: **Results of k-means clustering for tumor and tumor stroma separation**: above, the false color images produced according to pixel assignment to the respective clusters; below, cluster centroids and associated tissue type.

The resulting clusters and the false color images of the processed maps, where pixels are colored according to the corresponding cluster are reported in figure 4.26.
Comparing the clusters centroids, it is possible to see that cluster 2 features the lowest lipid-to-proteins ratio, followed by clusters 4 and then by 1, 5 and 3. Clusters 2 and 4 additionally appear to have their maximum intensity in correspondence to a wavenumber value slightly shifted on the right. This behaviour could be associated to a greater contribution to the signal in the CH region coming from DNA, since also a small bending in

correspondence to the 2954 cm^{-1} appears, usually associated to nucleic acids signal (see fig. 1.7 for comparison); however, attributing the change in shape of the CH region peak solely to a stronger DNA contribution is not entirely realistic. Such strong difference in DNA concentration is not so likely to be observed in tumor stroma, especially if compared to tumor tissue: tumor stroma in fact can be characterized by a higher cell density, but is more often constituted by extracellular matrix components, fibroblasts, blood vessel tissues originating from angiogenetic processes, which do not directly translate into a higher nuclei density. This is particularly true in the context of a comparison with tumor tissue, which is also known to be subject to rapid proliferation and multiplication of malignant cells and so to a quite crowded tissue environment. It is therefore more reasonable to interpret the feature of cluster 2 and cluster 4 as linked for its majority to a lower lipidic content and a greater protein presence, with some possible but minor contributions of increased DNA signal linked to a locally packed distribution of cells.

The greater presence of protein-related structures in tumor stroma was also verified by computing the average spectrum in the fingerprint region of all the pixels assigned to each cluster. Considering for example the results obtained for map 6c in figure 4.27, from those average spectra we see that also in the fingerprint region clusters associated to stroma (cluster 2 and 4) tend to have larger signal intensity for protein signal at 1460 cm^{-1} and amide I band (in particular at 1660 cm^{-1}); also peaks associated to collagen appear more prominently (932 cm^{-1} and 1068 cm^{-1}) and opposite ratios between signals at 1250 cm^{-1} (collagen) and 1314-1339 cm^{-1} (collagen, but also lipids and nucleic acids) can be noticed. The latters will be also explored more in depth in another analysis presented later. No peaks linked to DNA or nucleic acids signals seem to be visibly stronger with respect to clusters 1, 3 and 5, therefore making it reasonable to explain the different shape of the CH region peak with the lower lipid to protein content rather then an increased contribution from DNA.

By looking at the false color images in 4.28, it is first of all possible to see that with this choice of clustering it is possible to notice distinctions from different tissue types provided in the annotations. K-means clustering successfully allows to separate between tumor tissue and tumor stroma; indeed, map 8a can be used as a reference, and we can see from cluster assignment the same sharp distinction present in the annotations, with tumor tissue localized in the upper-left side of the image and the rest of the field of view covering tumor stroma (upper part of figure 4.28, third row). In particular, clusters 2 and 4 appear to be dominant in the latter tissue type.
On the other hand, map 6b is collected in a position where, according to annotation, almost only tumor tissue should be present; coherently with this and with the clusters

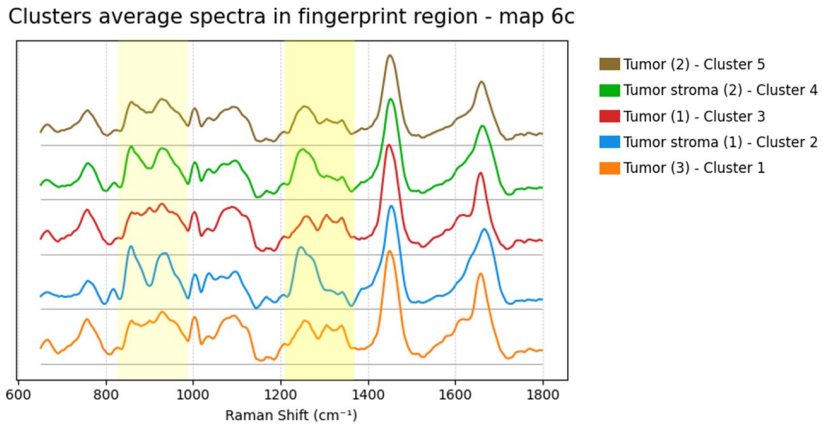


Figure 4.27: **Average spectra in fingerprint region of pixels associated to different clusters in map 6c**. The two band showing greatest differences from tumor-associated and tumor stroma-associated clusters are highlighted in yellow (932-1068 cm^{-1} and 1250-1314, 1339 cm^{-1}).

distribution inside map 8a, pixels of map 6b are almost entirely assigned to clusters 1, 3 and 5 (figure 4.26).
Moving finally to maps 6a and 6c, the field of view should be comprising, for its majority, tumor tissue; however, also some tumor stroma areas are included, with a distribution that is less easy to spot and separate, also considering that H&E annotations are provided on contiguous tissue slices, therefore not all structures have an exactly identical correspondence on the measured slices.
The colors provided by cluster assignment show that the patterns of cluster 2 and cluster 4 pixels on the Raman hyperspectral maps is very similar to the shape of tumor stroma classification provided by histopatologists. In figure 4.28 the false color image from k-means clustering, the brightfield 20x image of the mapped area, the false color image obtained with raw Raman data and the reference H&E annotated image are shown side by side, with yellow lines highlighting peculiar distributions of tumor stroma. Even if the local heterogeneity of the tissue can be already perceived from the brightfield image, k-means clustering provides a classification which is also supported by spectral data and relative biochemical information. It is also interesting to consider how this procedure, if properly tested and validated on a sufficiently large dataset, with maps belonging to different patients, has a great potentiality in aiding the process of manual annotation. False color images produced by raw data based, for example, only on intensity color scaling is not able to reproduce those structures, which instead is achievable with methods able to catch similar features distributed inside the dataset.

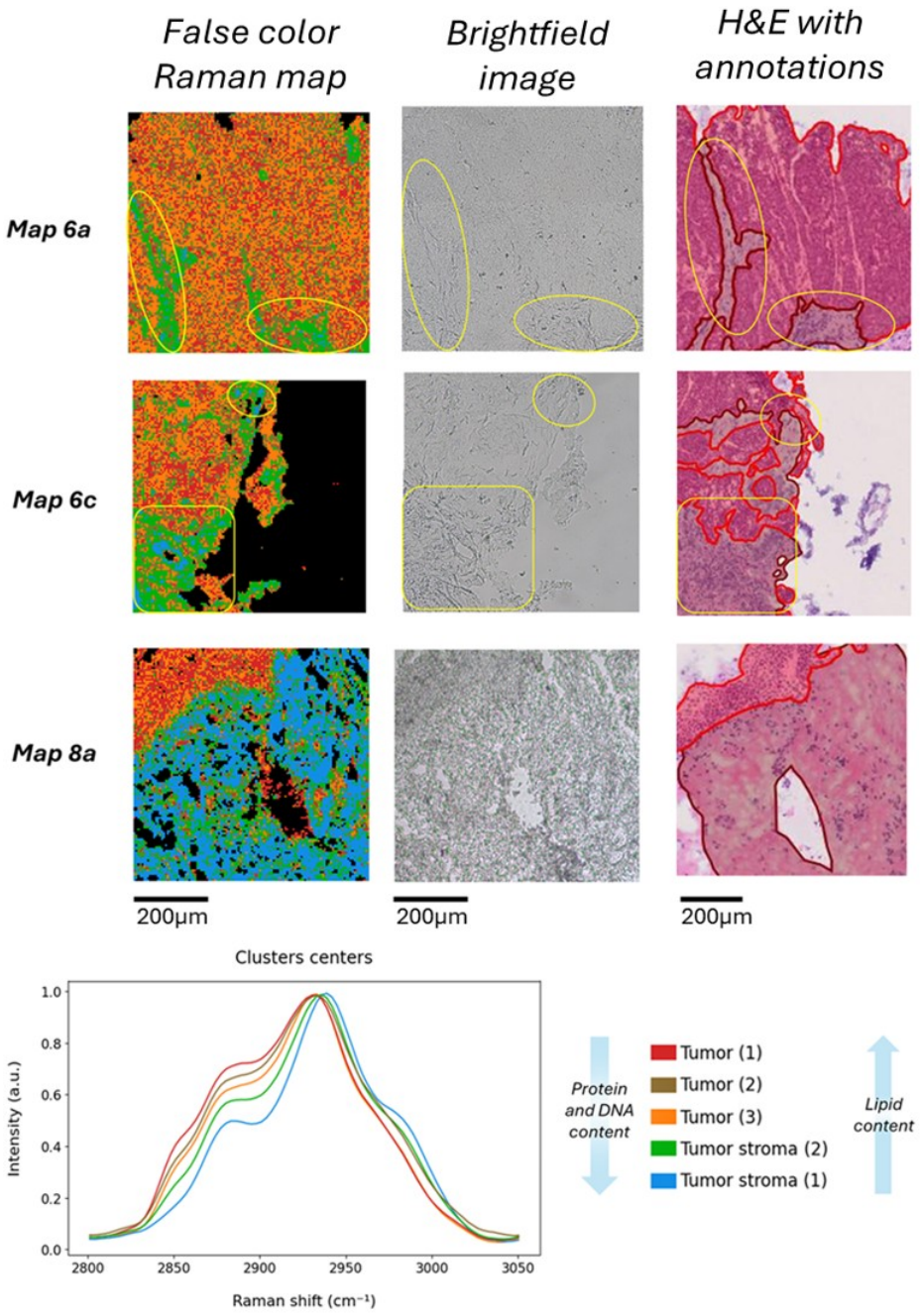


Figure 4.28: **Details on false color images of maps 6a, 6c and 8a**: from left to right, the false color Raman map produced with k-means clustering, the 20x brightfield image and the approximate corresponding area on the H&E annotations of the 3 maps featuring a mixture of tumor (red) and tumor stroma (dark red) tissue. The latter has been localized with yellow lines in maps 6a and 6c.

As it was mentioned previously, to have more details about the biochemical characterization of pixels associated to different clusters, the average spectrum in the fingerprint

region was also investigated. Among the mentioned peaks of interest suggesting that a greater collagen contribution is measured (fig. 4.27), attention has been captured by the changes in peak intensity ratios between the collagen signal at the wavenumber value of 1250 cm^{-1} and on the peak intensity ratio of the two adjacent band centered at 1314 cm^{-1} and 1339 cm^{-1}. According to [43], both of them can be again linked to collagen, but at the same time contributions to the signal can come from nucleic acids and lipids.

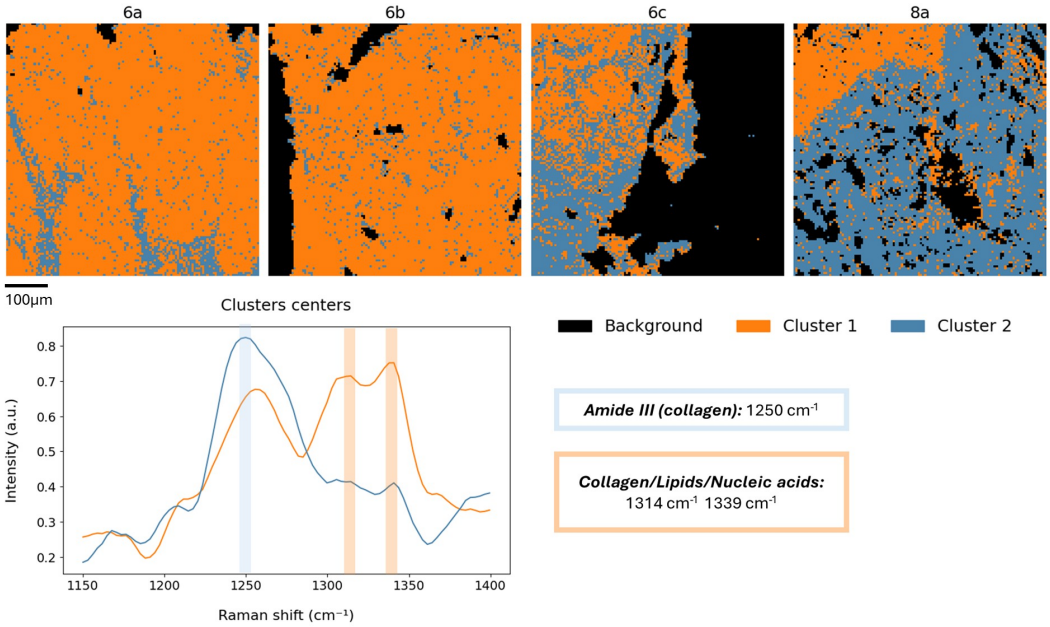


Figure 4.29: **Results of k-means clustering in 1150-1400 cm^{-1} spectral window**.

Since the difference in the ratio of the two peaks was quite evident and consistent, a k-means clustering limited in the spectral window 1150-1400 cm^{-1} was tested with 2 clusters to be identified and producing the output reported in fig. 4.29. The obtained centroids catch the differences carried by the different collagen to lipids/nucleic acids ratio: the first cluster has a features an higher collagen peak, while the second cluster shows the opposite situation.
Comparing the masks generated by the clusters produced in this trial with the ones previously obtained in the CH region analysis (fig. 4.26), quite good agreement is found in the assignment of clusters to the respective tissue type areas:

- Pixels assigned to cluster 2 (fig. 4.29, in blue), featuring high collagen content correspond to those previously assigned to clusters 2 and 4 (fig. 4.26) representing tumor stroma;
- Pixels assigned to cluster 1 (fig. 4.29, in orange), which on the contrary has higher

lipid and nucleic acids signal, correspond to those of clusters 1, 3 and 5 (fig. 4.26), associated to tumor tissue.

The results obtained with a 2-cluster approach in (1150-1400 cm^{-1}) provide a foundation, from the biochemical point of view, for the separation of pixels and tissue components provided by the algorithm. It is interesting to notice that, thanks to its greater specificity, the results obtained regarding map 6c appear to be even more accurate than those produced with a clustering on signals from the CH region, producing a division that appears more truthful with respect to the annotations and the H&E stained images (fig. 4.28). With appropriate testing and validation, an approach of this kind has indeed great potentiality for accurate labeling and classification of tissue structures; in order to achieve such result and validate this hypothesis, more accurate comparison with H&E staining and annotations on the very same tissue slice that being measured is mandatory, and examples from different patients, which are not available at the moment, is fundamental for a generalization of the method.

Obtaining a greater lipidic component in the tumor region with respect to the surrounding stroma can appear as an unusual conclusion, since wide evidence is present in literature stating an opposite trend in different cancer types, such as for breast cancer [45]. However, this may not be automatically translated to all the cases of HNSCC tumors; in particular, studies underline that the characteristics of tumor microenvironment can affect cancer cells lipid metabolism, leading to an increased production in case of lipid-depleted TME [46]. In addition, some cases are presented in literature reporting an altered lipid metabolism specifically for HNSCC and OSCC (oral squamous cell carcinoma), characterized in particular by an enhanced lipid biosyntesis in cancer cells [47], [48].
The main outcome from the experimental results in this thesis is generally more related to the significance of the collagen component for the tumor and tumor stroma distinction, whose higher presence in tumor stroma could also be linked to the presence of specific structures that promote its synthesis more than it happens in proximity of tumor cells.

Multi-patient skeletal muscle identification

The same approach just illustrated can be generalized also including maps from different patients and other tissue types besides tumor and tumor stroma. This is the case for the results shown in image 4.30 where maps 2d (patient SC8), 7b (patient SC7) and 51c, 51e and 53d (patient SC23) have been processed together.

Map name	Patient	Area	Timepoint	Treatment
2d	SC8	Sk. Muscle	day 0	non-treated
7b	SC7	mixed	day 3	non-treated
51c	SC23	mixed	day 0	non-treated
51e	SC23	mixed	day 0	non-treated

Table 4.7: **Dataset for multi-patient k-means clustering.**. For maps covering a region with heterogeneous tissue composition, the area classification is indicated as "mixed".

Again, k-means clustering in the CH region is able to separate quite well different classes (fig. 4.30), in this case tumor tissue, skeletal muscle tissue and lipid droplets, providing false color imaging that also introduce some complementary information to annotations on H&E slices. It is the case, for example, of map 51e (used in the example of the manual annotation tool in fig. 4.11). In particular, skeletal muscle and epithelial tissue (generically labeled as healthy) could be separated manually on the acquired Raman map thanks to the presence of a tissue structure that could be easily separated from the others even from the direct observation of the sample under the microscope. The bottom right side of the map, with transversally orientated tissue fibers, has been assigned to skeletal muscle class. However, k-means clustering allows to separate with even more detail, allowing the different biochemical composition to emerge: layers of more protein-rich tissue alternates with other healthy portions of tissue that instead are characterized by a composition more similar to that of the epithelial one.
The false color image of 4.30 is produced following the same procedure that has been illustrated before: after foreground pixels identification, k-means clustering has been applied in the CH region with a MinMax normalization.

The different tissue regions are assigned as follows:

- Cluster number 1 (fig. 4.30) shows similarities with clusters 2 and 4 seen in the previous example in fig. 4.28, with a low lipid to protein ratio, a small but non negligible contribution from DNA and a maximum located at an higher Raman shift with respect to all the other clusters. Pixels assigned to cluster 1 are once again those related to stroma and possible portions of tumor tissue (map 7b). Cluster 1 is also present in map 51e in correspondence to healthy tissues (general classification);
- Cluster number 3 and 4 show a lower DNA signal but still a lower lipid to protein ratio with respect to clusters 2 and 5. Pixels corresponding to skeletal muscle tissue are mostly assigned to these two clusters;

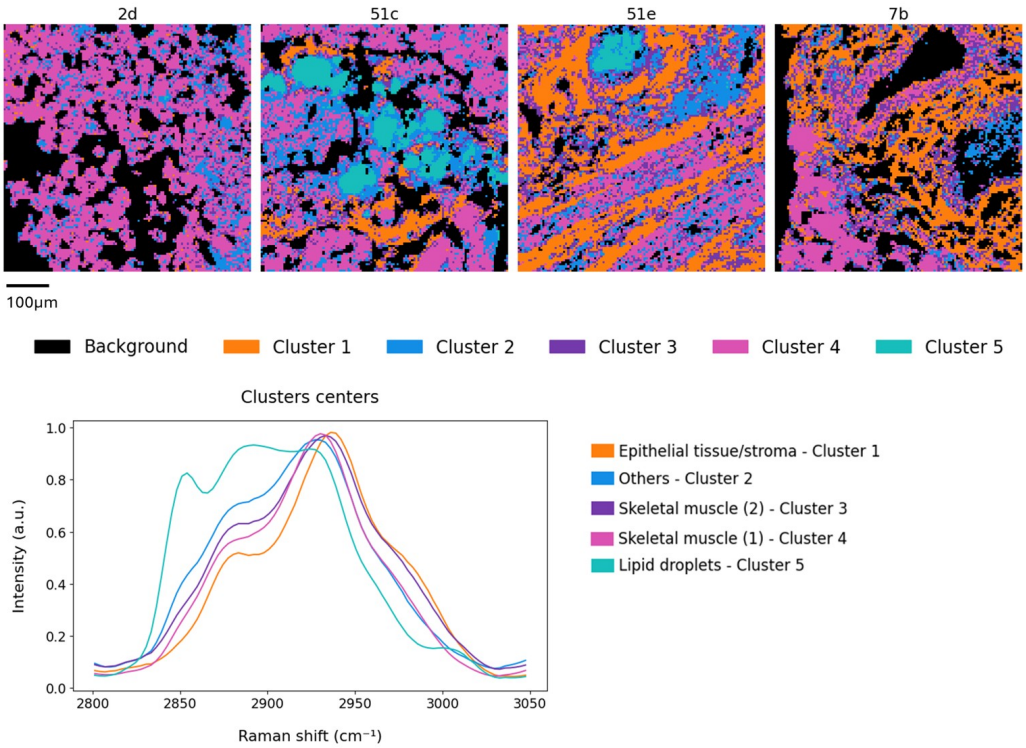


Figure 4.30: **Multi-patient k-means clustering in CH region**.

- Cluster 5 features a unique shape that separates it from the rest of the clusters centroids. The shape resembles with great clarity the one typical from lipid contribution to the CH region peak. This cluster effectively groups lipid droplets present in map 51c. The assignment of the portion of map 51e in cluster 5 will be the discussed;
- Cluster 2 features a discrete lipidic content, and reduced protein and DNA components. No specific tissue type has been identified for this cluster; from observation of the H&E contiguous slice, one possibility is to associated pixels belonging to this cluster to cytoplasm or connective tissue with low fibrous material.

The point of attention in this analysis concerns the assignment of pixels of map 51e to cluster number 5. In particular, as it can be noticed by observing brightfield images of the mapped areas reported in figure 4.31, the structures present in map 51c and map 51e are different one from the other: the first ones, on the bottom of 4.31, appear in fact with a well defined round or oval shape and a turbid-like color and can be identified as lipid droplets, which consist in small accumulations of fatty material. What can be seen in map 51e (top of 4.31) is instead very different at least by its looks from observation under the microscope. Such different appearance of these structures suggests that also

their biochemical composition is very likely to be different, as it is in fact verified by the two average spectra of the pixels of both maps assigned to cluster 5 and plotted in fig. 4.32.

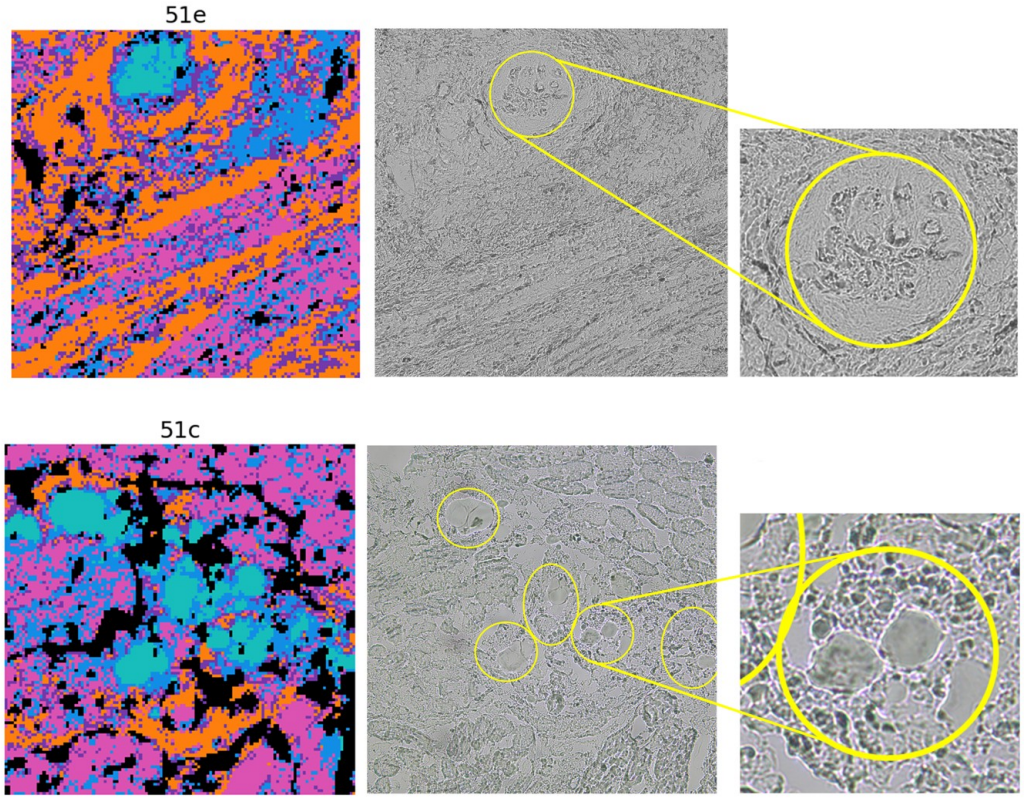


Figure 4.31: **Detail on tissue areas assigned to cluster number 5**: from the observation of brightfield images, the structures classified in cluster 5 for map 51e and 51c appear to be of two different kinds.

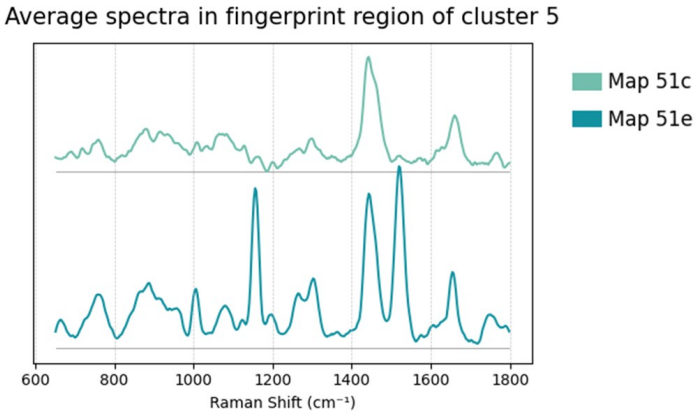


Figure 4.32: **Average spectrum in fingerprint region of pixels assigned to cluster 5**. Pixels belonging to map 51e show very peculiar features, characterized by high intensity and well defined peaks. Quite differently, map 51c shows prominent peaks in correspondence to amide I band (1660 cm^{-1}) and proteins (1453 cm^{-1}).

Performing again a k-means clustering, but focusing in this case in the fingerprint region only, we can see a distinction between the two tissue elements (figure 4.33). In order to avoid separations influenced by the effectiveness of the quartz substrate signal removal, the spectral region was limited to the range of 1100 - 1800 cm^{-1}. The normalization algorithm instead was chosen as L2 vector norm, which proved to be less affected by noise, providing better results with cluster centroids less affected by a vertical offset.

We see that the cluster assignment (fig. 4.31) resembles the one obtained with the CH region analysis (fig. 4.30). Though the separation of the different tissue structures is less sharp, the two elements of map 51c and 51e are now distinguished and associated to two different clusters. Leaving aside cluster number 5, it is possible to notice that Raman peaks corresponding to collagen in amide III and amide I bands (respectively at 1247 cm^{-1} and 1663 cm^{-1}) and the one related to proteins at 1253 cm^{-1} are evidently present in all clusters; the difference between one cluster and the other mostly lies in intensity differences and relative peak heights:

- Cluster 1 features the highest signal intensities, followed by cluster 4, 3 and lastly by cluster 2 , which probably tends to incorporate pixels with lower quality signal, as it can be seen from the more noisy and not well defined peaks shape of the cluster centroids;
- Cluster 4 presents a much higher signal in amide I region with respect to the other clusters, for which the peak at 1660 cm^{-1} is always smaller than the one of proteins

at 1253 cm^{-1};

- Clusters number 1 and 3 feature an higher ratio between signal of collagen at 1247 cm^{-1} and the signal at 1311 cm^{-1} and 1339 cm^{-1}, which can be still attributed to collagen but are also contributed by lipids and nucleic acids signals. This higher ratio could suggest a greater ratio between the presence of collagen and of other biological components, such as cells (responsible for the collection of DNA signal from their nuclei) or others constituted by fats. This has been a useful feature that allowed to easily separate tumor stroma from tumor tissue; indeed, this features seems to be shared also by healthy stroma (map 51e, cluster 2) and tumor stroma in map 7b.
- Cluster number 5 shows very peculiar spectral features, with very sharp and high intensity peaks at 1156 cm^{-1} and 1520 cm^{-1}. Such strong signals at these Raman shifts, which can be respectively assigned to the C-C and C=C bonds, are characteristic features of carotenoids. In particular, carotenoids are natural organic molecules that are soluble in lipids and are characterized by a linear, conjugated polyene chain, for which numerous C-C and C=C bonds are present within a single molecule; this results in very pronounced Raman peaks specifically for wavenumbers 1156 cm^{-1} and 1520 cm^{-1} [49]. The presence of carotenoids therefore prove to be the discriminant factor for the structures identified in maps 51e and 51c.

After further investigation and feedback from collaborators, it was finally verified that the structure seen in map 51e could be probably recognized as foam cells, a particular type of cells usually related to macrophages which get filled with modified low-density lipoproteins (LDL) and other lipids.
One final remark has to be done regarding the signal collected in lipid droplets in map 51c. Indeed, intensities of the Raman peaks associated to carotenoids are unusually high with respect to the average intensity collected from biomolecules. This effect can actually be linked to a phenomenon of resonance. Carotenoids, in fact, have an absorption spectrum which mainly covers the blue-green range of visible light, but also extends to higher wavelengths, with non-null values also at 660nm, that is the wavelength of the laser light employed for the excitation of the sample. By meeting the resonance condition, intensity gain of from 3 up to 6 orders of magnitude can be achieved [50], thus producing spectra similar to those of cluster 5.

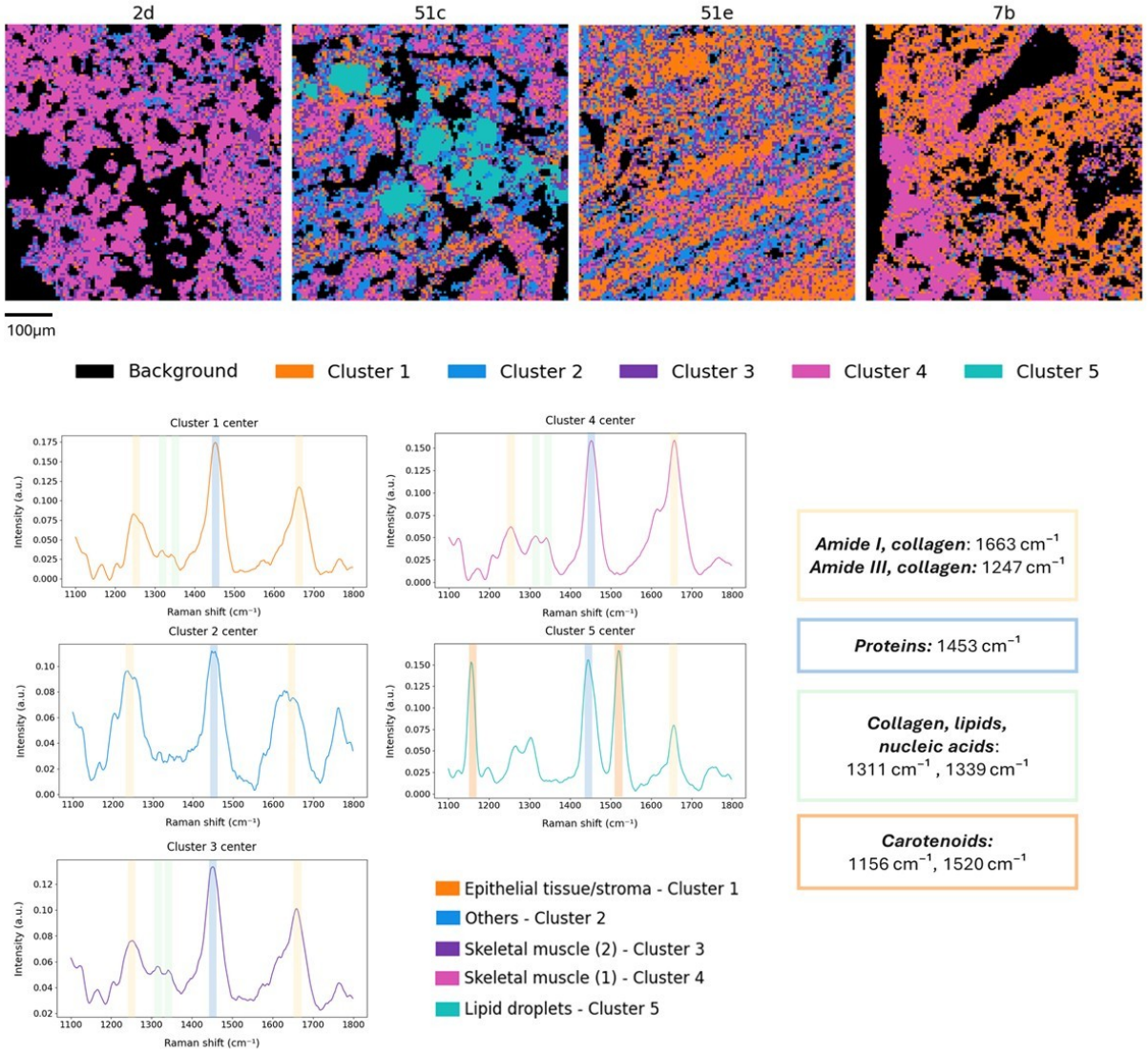


Figure 4.33: **Multi-patient k-means clustering in fingerprint region**.

From the two examples reported, we understand that a k-means clustering approach for tissue characterization is feasible also with a dataset comprising maps from different patients and including tissue classes not limited only to tumor stroma and tumor cells. The results produced by considering the CH region alone provides to be already quite effective in the identification of the different tissue classification within a set of maps, allowing insightful false-color images with sharp separations mostly based on varying lipid to protein ratios, but accounting also for different nucleic acids contributions. However, as it is known, the CH region suffers the limit of lacking the biochemical specificity which is instead proper of the fingerprint region: this leads to an imprecise classification which can only be overcome by moving to much higher clustering levels. The investigation of

the fingerprint region can instead lead to results which are more chemically specific, enabling a successful discrimination of the previously non distinguishable structures without increasing to much the number of clusters required, thus preserving also the clarity and imaging properties of the tool.

5 | Conclusions and future developments

The work presented in this thesis focused on the application of spontaneous Raman imaging of ex-vivo, patient-derived HNSCC slice cultures, in particular with the goal of achieving a biochemical characterization of the samples that could be linked to local tissue composition. The investigations performed proved the feasibility of discriminating different types of tissues, as well as the possibility of obtaining preliminary insights of the biochemical modifications undergoing with the passing of time with or without the application of a treatment during the cultivation.

From a methodological perspective, one of the main achievements was the design and implementation of a preprocessing pipeline tailored for a large dataset of Raman hyperspectral maps. This, in particular, combined standard procedures for spectral data preprocessing with custom-developed Python tools targeting specific needs of the dataset under analysis, such as the identification and masking of artifacts-related pixels, background signal subtraction and manual annotation on acquired Raman maps, the latter enabling a pixel-wise classification with respect to the involved different tissue types and allowing to account for intra-map heterogeneity.
Additionally, tools development included also the implementation of unsupervised, multivariate analysis algorithms (PCA and k-means clustering) to support the analysis across multiple maps, allowing for more comprehensive global analyses and extending the application of those algorithms beyond the single map level of the already available data analysis tools.

Experimental results confirmed that, with the methodology adopted for Raman spectroscopy data collection and preprocessing, it is possible to discriminate between tumor tissue, tumor stroma, skeletal muscle, and other healthy structures by making use of the k-means cluster algorithm, being in addition provided with a biochemical basis for these distinctions. The analysis highlighted the importance of selecting the appropriate spectral region of interest for the accuracy and soundness of the classification. In addition, the

ability to differentiate tissue types by their biochemical features supports the validity and robustness of the adopted methodology for measurements and preprocessing, an aspect that is crucial for advancing tissue characterization, studying its evolution, and investigating changes resulting from treatment. Moreover, this approach offers the potential to identify distinctive biochemical features that may eventually be linked to therapeutic responses.
Using principal component analysis, preliminary spectral differences observed in a single patient were generalized to a multi-patient analysis, offering early insights into the effects of PD-L1 protein application within four days of treatment. Notably, some patients exhibited an increase in protein and DNA content, suggesting an increase in cancer proliferation as a consequence of the suppressed immune response; at the same time, different trends were observed among patients with a specific HNSCC subtype (non-keratinizing oropharynx SCC in HPV-positive patients), a result remarking once more the complex nature of cancer leading to consequent variability across cases.

Considering the framework of this work, positioned in an early stage of a project, the results demonstrate a promising perspective for the application of Raman spectroscopy with the methodology and approaches employed so far. A criticality regarding the effective removal of background signal emerged, with suboptimal performances observed in a limited but non-negligible number of cases. Refinements and optimization of the substrate signal subtraction will be indeed one main objective for a further improvement in effectiveness and scalability of the preprocessing procedure.

For the moment, the analyses were focused only on the changes induced by immune system suppression by means of extrenally provided PD-L1 protein. Investigations regarding different therapeutic approaches, including immunotherapy, chemotherapy and radiotherapy are planned as next steps.
The dataset has been currently limited to four patients, on a small set of timepoints; indeed, the future availability of precise and specific information, such as H&E staining and annotations on the very same exact tissue slices measured, will be crucial for a full exploitation of the dataset and for more punctual and precise classifications, possibly with the introduction of tools enabling highly precision and automatic data coregistration.
The expansion of the dataset to a larger number of patients will be required for a generalization of the findings. Finally, the integration of additional data analysis tools will be fundamental, especially for more advanced supervised methods and machine learning techniques. This will allow the implementation of classification algorithms and evaluations based on discriminative power and diagnostic accuracy, further enhancing the analytical capabilities and insights deriving from data.

Bibliography

[1] Gábor Keresztury. Raman spectroscopy: Theory. In John M. Chalmers and Peter R. Griffiths, editors, *Handbook of Vibrational Spectroscopy*, volume 1, pages 71–87. John Wiley & Sons Ltd, 2002.

[2] D. Cialla-May, M. Schmitt, , and J. Popp. Theoretical principles of raman spectroscopy. *Physical Sciences Reviews*, 4(6):20170040, 2019. doi: doi:10.1515/psr-2017-0040.

[3] Jiabao Xu, Tong Yu, Christos E. Zois, Ji-Xin Cheng, Yuguo Tang, Adrian L. Harris, and Wei E. Huang. Unveiling cancer metabolism through spontaneous and coherent raman spectroscopy and stable isotope probing. *Cancers*, 13(7):1718, 2021.

[4] Robin R. Jones, David C. Hooper, Liwu Zhang, Daniel Wolverson, and Ventsislav K. Valev. Raman techniques: Fundamentals and frontiers. *Nanoscale Research Letters*, 14(231), 2019. doi: 10.1186/s11671-019-3039-2.

[5] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74:229–263, 2024. doi: 10.3322/caac.21834. Accessed: 2024-07-08.

[6] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. W. W. Norton & Company, 7 edition, 2022.

[7] Roberta Lugano, Mohanraj Ramachandran, and Anna Dimberg. Tumor angiogenesis: causes, consequences, challenges and opportunities. *Cellular and Molecular Life Sciences*, 77(9):1745–1770, 2020. doi: 10.1007/s00018-019-03351-7.

[8] Uttpal Anand, Abhijit Dey, Arvind K. Singh Chandel, Rupa Sanyal, Amarnath Mishra, Devendra Kumar Pandey, Valentina De Falco, Arun Upadhyay, Ramesh Kandimalla, Anupama Chaudhary, Jaspreet Kaur Dhanjal, Saikat Dewanjee, Jayalakshmi Vallamkondu, and Jose M. Perez de la Lastra. Cancer chemotherapy and be-

yond: Current status, drug candidates, associated risks and progress in targeted therapeutics. *Genes & Diseases*, 10:1367–1401, 2023. doi: 10.1016/j.gendis.2022.02.007.

[9] Noriko Mitsuiki, Charlotte Schwab, and Bodo Grimbacher. What did we learn from CTLA-4 insufficiency on the human immune system? *Immunological Reviews*, 287 (1):33–49, 2019. doi: 10.1111/imr.12721.

[10] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh. Cancer and radiation therapy: Current advances and future directions. *International Journal of Medical Sciences*, 9(3):193–199, 2012. doi: 10.7150/ijms.3635.

[11] Elham Bidram, Yasaman Esmaeili, Hadi Ranji-Burachaloo, Nuha Al-Zaubaid, Ali Zarrabi, Alastair Stewart, and Dave E. Dunstan. A concise review on cancer treatment methods and delivery systems. *Journal of Drug Delivery Science and Technology*, 54:101350, 2019. doi: 10.1016/j.jddst.2019.101350.

[12] Geoffrey T. Gibney, Louis M. Weiner, and Michael B. Atkins. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncology*, 17(12):e542–e551, 2016. doi: 10.1016/S1470-2045(16)30406-5.

[13] Santosh K. Paidi, Paola Monterroso Diaz, Sina Dadgar, Samir V. Jenkins, Charles M. Quick, Robert J. Griffin, Ruud P.M. Dings, Narasimhan Rajaram, and Ishan Barman. Label-free raman spectroscopy reveals signatures of radiation resistance in the tumor microenvironment. *Cancer Research*, 79(8):2054–2064, 2019. doi: 10.1158/0008-5472.CAN-18-2732.

[14] Christoph Krafft, Michael Schmitt, Iwan W. Schie, Dana Cialla-May, Christian Matthäus, Thomas Bocklitz, and Jürgen Popp. Label-free molecular imaging of biological cells and tissues by linear and nonlinear raman spectroscopic approaches. *Angewandte Chemie International Edition*, 56(16):4392–4430, 2017. PMID: 28027574.

[15] Santosh Kumar Paidi, Joel Rodriguez Troncoso, Piyush Raj, Paola Monterroso Diaz, Jesse D. Ivers, David E. Lee, Nathan L. Avaritt, Allen J. Gies, Charles M. Quick, Stephanie D. Byrum, Alan J. Tackett, Narasimhan Rajaram, and Ishan Barman. Raman spectroscopy and machine learning reveals early tumor microenvironmental changes induced by immunotherapy. *Cancer Research*, 81(22):5745–5755, 2021. doi: 10.1158/0008-5472.CAN-21-1438.

[16] Varsha Karunakaran, Sina Dadgar, Santosh K. Paidi, April F. Mordi, Whitney A. Lowe, Umme Marium Mim, Jesse D. Ivers, Joel I. Rodriguez Troncoso, Jared A. McPeake, Alric Fernandes, Sanidhya D. Tripathi, Ishan Barman, and Narasimhan

Rajaram. Investigating in vivo tumor biomolecular changes following radiation therapy using raman spectroscopy. *ACS Omega*, 9(42):43025–43033, 2024. doi: 10.1021/acsomega.4c06096.

[17] Daniel E. Johnson, Barbara Burtness, C. René Leemans, Vivian Wai Yan Lui, Julie E. Bauman, and Jennifer R. Grandis. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*, 6(1):92, 2020. doi: 10.1038/s41572-020-00224-3.

[18] Rebecca D. Chernock, Samir K. El-Mofty, Wade L. Thorstad, Curtis A. Parvin, and James S. Lewis Jr. HPV-related nonkeratinizing squamous cell carcinoma of the oropharynx: Utility of microscopic features in predicting patient outcome. *Head and Neck Pathology*, 3(3):186–194, 2009. doi: 10.1007/s12105-009-0126-1.

[19] F. L. J. Cals, T. C. Bakker Schut, P. J. Caspers, R. J. Baatenburg de Jong, S. Koljenović, and G. J. Puppels. Raman spectroscopic analysis of the molecular composition of oral cavity squamous cell carcinoma and healthy tongue tissue. *Analyst*, 143(19): 4090–4102, 2018. doi: 10.1039/c7an02106b.

[20] Roberto Valdés, Stefan Stefanov, Stefano Chiussi, Miriam López-Alvarez, and Pío González. Pilot research on the evaluation and detection of head and neck squamous cell carcinoma by raman spectroscopy. *Journal of Raman Spectroscopy*, 45(7):550–557, 2014. doi: 10.1002/jrs.4498.

[21] F. L. Cals et al. Development and validation of raman spectroscopic classification models to discriminate tongue squamous cell carcinoma from non-tumorous tissue. *Oral Oncol.*, 60:41–47, 2016.

[22] Cornelia van Lanschot, Tom Bakker Schut, Elisa Barroso, Aniel Sewnaik, Jose Hardillo, Dominiek Monserez, Cees Meeuwis, Stijn Keereweer, Rob Baatenburg de Jong, Gerwin Puppels, and Senada Koljenović. Raman spectroscopy to discriminate laryngeal squamous cell carcinoma from non-cancerous surrounding tissue. *Lasers in Medical Science*, 38:193, 2023. doi: 10.1007/s10103-023-03849-4.

[23] Ola Ibrahim, Mary Toner, Stephen Flint, Hugh J Byrne, and Fiona M Lyng. The potential of raman spectroscopy in the diagnosis of dysplastic and malignant oral lesions. *Cancers*, 13(4):619, 2021. doi: 10.3390/cancers13040619.

[24] Bowen Yang, Xiaobo Dai, Zhixin Li, Zhenxin Wu, Shuai Chen, Chunjie Li, and Bing Yan. Noninvasive surface-enhanced raman spectroscopy outperforms combined positive score in predicting sensitivity to neoadjuvant immunotherapy in head and neck squamous cell carcinoma. *Oral Oncology*, 159:107105, 2024. doi: 10.1016/j.oraloncology.2024.107105.

[25] Bowen Yang, Xiaobo Dai, Zhixin Li, Shuai Chen, Chunjie Li, and Bing Yan. Application of surface-enhanced raman spectroscopy in head and neck cancer diagnosis. *ACS Publications*, 2025. doi: 10.1021/acs.analchem.4c02796.

[26] Yining Zhang, Zhenfang Li, Chengchi Zhang, Chengying Shao, Yanting Duan, Guowan Zheng, Yu Cai, Minghua Ge, and Jiajie Xu. Recent advances of photodiagnosis and treatment for head and neck squamous cell carcinoma. *Neoplasia*, 60: 101118, 2024. doi: 10.1016/j.neo.2024.101118.

[27] Adil Parvez, Furqan Choudhary, Priyal Mudgal, Rahila Khan, Kamal A. Qureshi, Humaira Farooqi, and Ashok Aspatwar. PD-1 and PD-L1: architects of immune symphony and immunotherapy breakthroughs in cancer treatment. *Frontiers in Immunology*, 14:1296341, 2023. doi: 10.3389/fimmu.2023.1296341.

[28] Andrea M.P. Romani. Cisplatin in cancer treatment. *Biochemical Pharmacology*, 206:115323, 2022. doi: 10.1016/j.bcp.2022.115323.

[29] Thomas W. Bocklitz, Shuxia Guo, Oleg Ryabchykov, Nadine Vogler, and Jürgen Popp. Raman based molecular imaging and analytics: A magic bullet for biomedical applications? *Analytical Chemistry*, 88(1):133–151, 2016. doi: 10.1021/acs.analchem.5b04665.

[30] Jan Gerretzen, Ewa Szymańska, Jeroen J. Jansen, Jacob Bart, Henk-Jan van Manen, Edwin R. van den Heuvel, and Lutgarde M. C. Buydens. Simple and effective way for data preprocessing selection based on design of experiments. *Analytical Chemistry*, 87(20):12096–12103, 2015. doi: 10.1021/acs.analchem.5b02832.

[31] Ramapp. URL `https://ramapp.io`.

[32] Z. Movasaghi, S. Rehman, and I. U. Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007. doi: 10.1080/05704920701410098.

[33] W. Fortunato, A. J. Chiquito, J. C. Galzerani, and J. R. Moro. Crystalline quality and phase purity of CVD diamond films studied by Raman spectroscopy. *Journal of Materials Science*, 42:7331–7336, 2007. doi: 10.1007/s10853-007-1575-0.

[34] Christoph Krafft, Stephan B. Sobottka, Gabriele Schackert, and Reiner Salzer. Near infrared raman spectroscopic mapping of native brain tissue and intracranial tumors. *The Analyst*, 130:1070–1077, 2005. doi: 10.1039/b419232j.

[35] F.-K. Lu, S. Basu, V. Igras, M. P. Hoang, M. Ji, D. Fu, G. R. Holtom, V. A. Neel, C. W. Freudiger, D. E. Fisher, and X. S. Xie. Label-free DNA imaging in vivo with

stimulated raman scattering microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37):11624–11629, 2015. doi: 10.1073/pnas.1515121112.

[36] Hugh J. Byrne, Peter Knief, Mark E. Keating, and Franck Bonnier. Spectral pre and post processing for infrared and raman spectroscopy of biological tissues and cells. *Chemical Society Reviews*, 45(8):1865–1878, 2016. doi: 10.1039/c5cs00440c.

[37] Ruihao Luo, Juergen Popp, and Thomas Bocklitz. Deep learning for raman spectroscopy: A review. *Analytica*, 3(3):287–301, 2022. doi: 10.3390/analytica3030020. Open Access under CC BY 4.0 license.

[38] Marco Ventura. Development and optimization of raman microscopy techniques for the pre-clinical study of osteopetrosis. Master's thesis, Politecnico di Milano, 2022.

[39] World Medical Association. World medical association declaration of helsinki: ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194, November 2013. doi: 10.1001/jama.2013.281053.

[40] Feng Zhang, Xiao-Jun Tang, Angxin Tong, Bin Wang, Jingwei Wang, Yangyu Lv, Chunrui Tang, and Jie Wang. Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method. *Spectroscopy Letters*, 53:1–12, 02 2020. doi: 10.1080/00387010.2020.1730908.

[41] Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

[42] Brooke D. Beier and Andrew J. Berger. Method for automated background subtraction from raman spectra containing known contaminants. *Analyst*, 134(6):1198–1202, 2009. doi: 10.1039/b821856k.

[43] Abdullah Chandra Sekhar Talari, Zanyar Movasaghi, Shazza Rehman, and Ihtesham ur Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 50(1):46–111, 2015. doi: 10.1080/05704928.2014.923902.

[44] Samir K. El-Mofty and Sushama Patil. Human papillomavirus (HPV)-related oropharyngeal nonkeratinizing squamous cell carcinoma: characterization of a distinct phenotype. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 101(3):339–345, 2006. doi: 10.1016/j.tripleo.2005.08.001. PMID: 16504868.

[45] Abigail S. Haka, Kathleen E. Shafer-Peltier, Mary Fitzmaurice, John Crowe, Ramachandra R. Dasari, and Michael S. Feld. Diagnosing breast cancer by using raman

spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12371–12376, August 2005. doi: 10.1073/pnas.0501390102.

[46] Patrick B. Jonker and Alexander Muir. Metabolic ripple effects – deciphering how lipid metabolism in cancer interfaces with the tumor microenvironment. *Disease Models & Mechanisms*, 17(9):dmm050814, 2024. doi: 10.1242/dmm.050814.

[47] P. Liu, Y. Wang, X. Li, Z. Liu, Y. Sun, H. Liu, Z. Shao, E. Jiang, X. Zhou, and Z. Shang. Enhanced lipid biosynthesis in oral squamous cell carcinoma cancer-associated fibroblasts contributes to tumor progression: Role of IL8/AKT/p-ACLY axis. *Cancer Science*, 115(5):1433–1445, 2024. doi: 10.1111/cas.1433.

[48] Janos Schmidt, Béla Kajtár, Kata Juhász, Mária Péter, Tamás Járai, András Burián, László Kereskai, Imre Gerlinger, Tamás Tornóczki, Gábor Balogh, László Vígh, László Márk, and Zsolt Balogi. Lipid and protein tumor markers for head and neck squamous cell carcinoma identified by imaging mass spectrometry. *Oncotarget*, 11 (28):2702–2717, 2020. doi: 10.18632/oncotarget.27615. Open Access under CC BY 3.0.

[49] Mindaugas Macernis, Denise Galzerano, Juozas Sulskus, Elizabeth Kish, Young-Hun Kim, Sangho Koo, Leonas Valkunas, and Bruno Robert. Resonance raman spectra of carotenoid molecules: Influence of methyl substitutions. *The Journal of Physical Chemistry A*, 119(1):56–66, 2015. doi: 10.1021/jp510426m.

[50] Jean Claude Merlin. Resonance raman spectroscopy of carotenoids and carotenoid-containing systems. *Pure & Applied Chemistry*, 57(5):785–792, 1985.

A | Appendix - Map nomenclature and dataset for the analysis

Each collected Raman map contains spectral information on a specific field of view inside a tissue slice; it is therefore essential to keep track of all the information relative to the mapped region, both in terms of tissue type classification (tumor, healthy, mixed...) and slice-related information, such as patient of belonging, timepoint, treatment condition. In order to simplify the management and storage of data, it was decided to adopt a nomenclature for which each map has an associated name; all the information abovementioned relative to the map, together with deeper details related to the corresponding patient clinical condition, are then summarized in a separate matrix. Python codes performing analyses retrieve the relative map wise (then translated to pixel wise) information by accessing the data matrix and filtering according to the map name.

For sake of simplicity, considering the large dimensions of the dataset, the names have been chosen as a combination of one number and one letter. In particular:

- When measuring a new tissue slice, with at least one difference in terms of treatment condition, timepoint replicate or patient, a new number is selected
- Multiple maps collected on the same tissue slice all share the same number, but change the letter part of the name
- The order of the assignment of numbers and letters follows the chronological order of acquisition. To preserve clarity, samples are measured in timepoint order (from t0 to the highest available); before moving from one timepoint to the following, all treatment conditions are measured The same goes when moving from one patient to the other: all timepoints, treatment conditions and replicates of a single patient are sampled before moving to the following one.
- Maps collected on replicates (provided for patient SC15 and SC17) are named with

the same number and different letter, since same patient, timepoint, treatment condition and H&E annotated slice are shared.

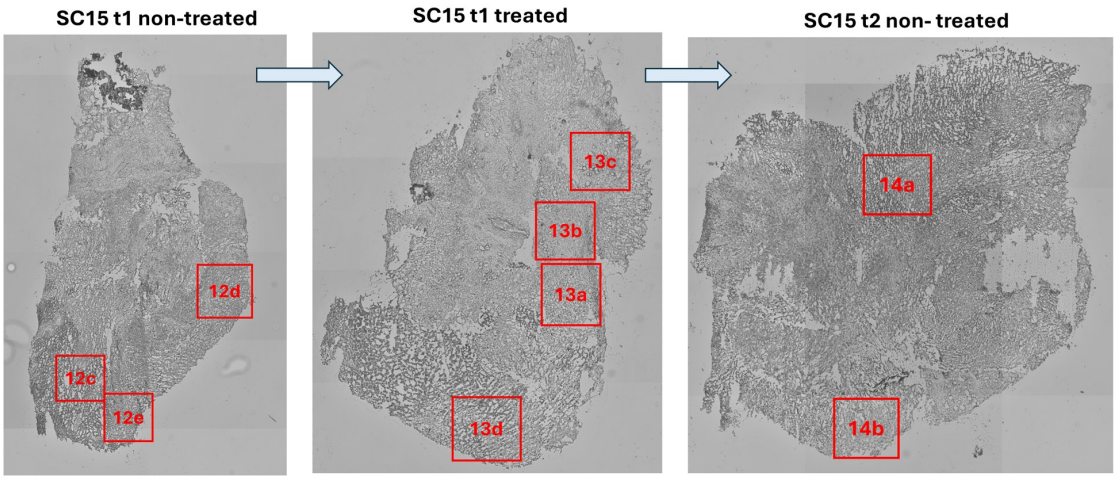


Figure A.1: **Nomenclature for different maps**: within the same slice, map names share the same number and are distinguished by the letter, while from one slice to another the number is changed. Arrows indicate the chronological order of map acquisition on different slices.

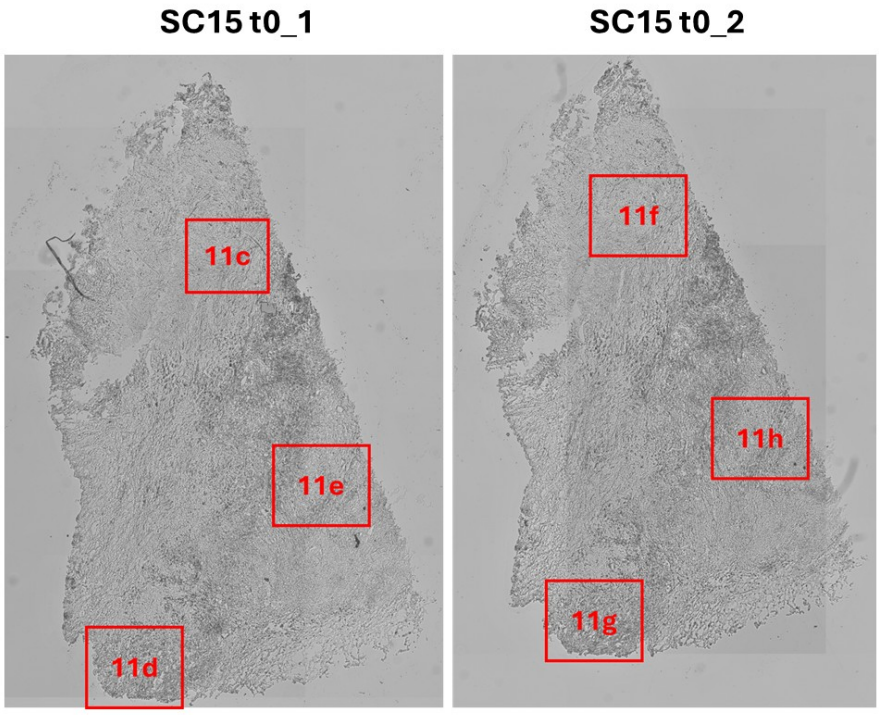


Figure A.2: **Nomenclature for replicates**: maps on two different tissue slices that represent biological replicates (_1, _2) still share the same number in the name.

List of Figures

1.1 Schematic representation of the possible scattering interactions 6
1.2 Angular dependence of the emitted radiant intensity of an oscillating dipole 7
1.3 The three spectral regions (fingerprint, silent and CH) on a sample Raman spectrum (toluene). 11
1.4 Schematic example of how cancer cells protect themselves from immune response 15
1.5 Locations of Head and Neck tumors [17] 20
1.6 Two examples of improper baseline correction 24
1.7 Lipid, protein and DNA contributions to CH region peak 26

3.1 Schematic representation of the experimental setup 31
3.2 Steps of the sample preparation procedure 34
3.3 Procedure to select regions to be measured 36
3.4 Measured Raman spectra of ArHg lamp (a) and toluene (b) used for wavenumber axis calibration 40
3.5 Raman spectrum of the white lamp: measured (a) and theoretical (b) . . . 41
3.6 Example of signal from cosmic rays 42
3.7 Raman spectrum measured inside quartz substrate 43

4.1 Sample Raman spectrum from map 6c 49
4.2 Step 3 of the pipeline: spectral crop 49
4.3 Step 4 of the pipeline: cosmic rays signal identification and correction . . . 50
4.4 Step 1 of the pipeline: spectral crop 50
4.5 Step 7 of the pipeline: normalization with L2 norm 51
4.6 Step 5 of the pipeline: spectral smoothing 52
4.7 Comparison between three different preprocessing pipelines 54
4.8 Example of EMSC for background signal subtraction 59
4.9 Example of Raman spectrum in a burnt spot 60
4.10 Example of artifacts masks for map 52i (patient SC23, day 2, Cisplatin 3,33 μM, mixed areas) 62

4.11 User interface for manual annotations on Raman maps 64
4.12 Schematic summary of acquisiton and pre-analysis steps 66
4.13 Examples of maps temporarily excluded from the analyses. 68
4.14 Schematic representation of the foreground identification process 71
4.15 Averages between different areas of map 7b 73
4.16 Differences between average spectra of tumoral tissue, skeletal muscle and blood vessel tissue in map 7b 74
4.17 Average spectrum of SC8 tumor tissue maps. 76
4.18 Average spectrum differences of treated and non-treated tumor tissue maps in patient SC8 77
4.19 Principal components 1, 6 and 8 81
4.20 Principal components 1, 2 and 3 (CH region) 81
4.21 Boxplots of PC1, PC6 and PC8 scores on full spectrum PCA analysis . . . 83
4.22 Boxplots of PC1, PC2 and PC3 scores of CH region PCA analysis 83
4.23 Principal components 1 and 2 in CH region 85
4.24 Boxplots of PC1 and PC2 scores 86
4.25 Positions of maps 6a, 6b, 6c and 8a 88
4.26 Results of k-means clustering for tumor and tumor stroma separation . . . 89
4.27 Average spectra in fingerprint region of pixels associated to different clusters in map 6c 91
4.28 Details on false color images of maps 6a, 6c and 8a 92
4.29 Results of k-means clustering in 1150-1400 cm^{-1} spectral window 93
4.30 Multi-patient k-means clustering in fingerprint region 96
4.31 Detail on tissue areas assigned to cluster number 5 97
4.32 Average spectrum in fingerprint region of pixels assigned to cluster 5 98
4.33 Multi-patient k-means clustering in fingerprint region 100

A.1 Nomenclature for different maps 112
A.2 Nomenclature for replicates 112

List of Tables

3.1 Specifications of the objective employed for measurements (MPLFLN50X) 33
3.2 Dataset information: available timepoints (a), treatments (b) and tissue slices number (c) for each patient. 37

4.1 Average SNR of single pixels in each map 56
4.2 SNR of the mean spectrum of foreground pixels for each map 56
4.3 Final dataset for analyses 69
4.4 Dataset for PCA analysis on SC8 maps 80
4.5 Dataset for PCA for PD-L1 treatment response on 4 patients 84
4.6 Dataset for tumor and tumor stroma K-means clustering 88
4.7 Dataset for multi-patient k-means clustering 95