SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Physics

PhD program in Physics and Astronomy Cycle XXXV

Curriculum in Plasma Physics and Biophysics

NOVEL HYBRID AI-PHASOR BASED TOOLS FOR REMOTE SENSING AND HISTOPATHOLOGICAL APPLICATIONS

Surname Scodellaro Name Riccardo

Registration number 769386

Tutor: Prof. Laura Sironi

Supervisor: Prof. Giuseppe Chirico

Coordinator: Prof. Stefano Ragazzi

ACADEMIC YEAR 2022/2023

Legend

Introduction

The phasor approach is a widely exploited mathematical tool which takes advantage of the Discrete Fourier Transform to decompose a signal into a sum of elementary functions, characterized by different frequencies. In the standard approach, the entire signal is projected into a point in a 2-dimensional space, named phasor plot, whose coordinates represent the real and imaginary part of the DFT and are strictly related to the signal properties and shape. Therefore, signals characterized by similar parameters lie closely within the phasor plot. Since it offers single point representations of complex signals, the phasor approach reduces data dimension, simplifying their handling and interpretation.

Moreover, the phasor approach has demonstrated to be a really efficient tool in order to analyze different experimental signals related to the biophysics field, as fluorescence lifetime (Digman et al. 2008), Second Harmonic Generation (Radaelli et al. 2017) and holography (Luo et al. 2016). However, despite this wide range of applications, no studies have coupled the phasor approach with artificial intelligence techniques to estimate physical parameters describing the functional form of the acquired signal.

In this scenario, during my PhD project, I coupled the standard phasor approach with supervised machine learning (ML) techniques, developing novel versatile methods able to analyze spectra and images in two different research fields, i.e. histopathology and remote sensing. In particular, this thesis proposes three different algorithms which couple the phasor approach to ML and statistics-based parameters retrieval pipelines: i-φ-MaLe, SuperμMAPPS and Φ-Norm.

i-φ-MaLe is specifically aimed to analyze the reflectance spectra of vegetation acquired by means of hyperspectral sensors. Here, the DFT is no longer computed on the entire spectrum, but the phasor approach has been applied on consecutive spectral windows to extract a set of parameters, assisted by a ML and a statistics-based retrieval algorithm. The relevance and impact of this study is the capability of i-φ-MaLe to simultaneously retrieve,

from the acquired signal, the top of canopy fluorescence spectrum F, the F spectrum corrected for leaf/canopy reabsorption (i.e. at photosystem level) F_{RC}, the quantum efficiency (Fqe) and three canopy-related biophysical parameters (Leaf Area Index, Chlorophyll content and the Absorbed Photosynthetically Active Radiation) in few milliseconds. Moreover, i-φ-MaLe is the first developed algorithm providing a consistent estimate of the F_{RC} and Fqe parameters from experimentally measurements acquired onsite: this is a crucial novelty in the Remote Sensing field since the analysis of F_{RC} and Fqe lead to new insights regarding the vegetation stress detection.

SuperμMAPPS analyzes the Polarization-dependent Second Harmonic Generation (P-SHG) signal of collagen to extract information related to its microarchitecture in terms of fibrils orientation angle and fibrils local disorder. The proposed tool extends the i-φ-MaLe algorithm structure by applying the windowed phasor approach to different permutations of the original P-SHG spectrum, allowing a fully automated estimate of the collagen microstructure with a significant reduction in data acquisition time. This will pave the way to an extensive characterization of the collagen properties to improve cancer early diagnosis by means of label-free imaging.

Φ-Norm automatically performs stain normalization of biopsy images acquired in different experimental conditions (i.e. different laboratories) by means of whole-slide scanners. In particular, Φ-Norm computes the DFT of the RGB color spectrum characterizing each pixel in an acquired image. Then, a geometrical-based pipeline applied in the phasor space maps the colors of any acquired image to a reference. Stain normalization tools are particularly relevant in the histopathology field since they reduce the discrepancies in the color content of the same biological structure visualized in different images, leading to an increased accuracy of AI algorithms applied for segmentation tasks.

The first part of the thesis provides a brief introduction regarding the standard phasor approach (**Section 1**) and the problems faced by the three algorithms: the deconvolution of fluorescence spectra and biophysical parameters related to the vegetation health from the

apparent reflectance spectra acquired at top of canopy (**Section 2**), the crucial role of stain normalization in the Digital Pathology field (**Section 3**) and the analysis of the collagen micro-architecture by exploiting the PSHG signal (**Section 4**).

The second part of the thesis concerns the description and validation of the algorithms and their application to real experimental scenarios. In particular, i-φ-MaLe has been at first validated by exploiting synthetic apparent reflectance spectra, considering different instrumental noise and algorithm tuning parameters (spectral windows width, number of exploited phasor planes, dataset size, etc.). Then, it has been applied to analyze real data acquired by spectrometers placed both at top of canopy and tower level (~100 m between the instrument and the canopy) in order to retrieve seasonal and daily trends characterizing the fluorescence and the involved biophysical variables (**Section 5**). Φ-Norm has been applied to liver sections images stained with Hematoxylin and Eosin (H&E) or PicroSirius Red (SR) and acquired by means of whole-slide scanners. The proposed stain normalization tool has been quantitatively compared with the results obtained through other already published methods in terms of color similarity of the same biological structures visualized in different images (**Section 6**). SuperμMAPPS accuracy has been quantified and compared with standard fitting approaches on synthetic and experimental P-SHG images of murine tendons and human tissues. Moreover, the collagen-related features provided by SuperμMAPPS have been exploited to detect early stage breast tumor in human biopsies (**Section 7**).

The final part of the manuscript provides a brief discussion of the results achieved during my PhD project and some perspectives concerning the future developments of each algorithm (**Section 8**).

1. Phasor Approach & Machine Learning

In this section, a brief introduction to the most relevant theoretical aspects and properties involving the phasor approach is provided, together with some basic principles of the Supervised Machine Learning. Indeed, this information is essential to clearly understand the three proposed pipelines (i-φ-MaLe, SuperμMAPPS and Φ-Norm), which are based on the coupling of machine learning with the phasor approach.

1.1 Discrete Fourier Transform

The Fourier Transform (FT) is a widely exploited computational technique able to decompose a signal into a sum of elementary functions, which are easier to analyze (Chapeau-Blondeau et al. 2020). These novel signals are complex, periodic, and characterized by different frequencies: these properties allow to exploit the FT to analyze the original signal in terms of phase or amplitude. Indeed, thanks to the FT application, the entire noise contribution affecting a signal at a defined frequency can be deconvolved and deleted without losing information concerning other frequency domains (Schoukens et al. 1986). At the state of the art, the FT is widely exploited among the scientific community and plays a key-role in different fields, from acoustic (Salah et al. 2021) and cryptography (Yang et al. 2014) to artifact reduction (Reddy et al. 2008) and data compression (Kaushik et al. 2014). Moreover, novel properties and methods able to reduce its computational cost have been recently developed (Rajabi et al. 2022, Burova 2021).

Briefly, the FT X(f), expressed in the frequency domain, of a signal x(t), in the time domain, can be defined as:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-2\pi i t f} dt \qquad (1)$$

Unfortunately, in experimental scenarios, each signal is not continuous, but it is characterized by a defined number N of acquired points. As a consequence, **Eq. 1** has to be adapted to discrete signals, obtaining the Discrete Fourier Transform (DFT):

$$X(k) = \sum_{n=1}^{N} \quad x(n)e^{\frac{2\pi ikn}{N}} \qquad (2)$$

In particular, $0 \leq k \leq N$, where k is the harmonic number. Similarly to the FT output, *X(k)* encodes the amplitude and phase of a sinusoidal wave of frequency $f = \frac{k}{n}$ since it exploits the Euler's formula:

$$e^{\frac{2\pi jkn}{N}} = cos\left(\frac{2\pi kn}{N}\right) + i\ sin\ (\frac{2\pi kn}{N}) \qquad (3)$$

Since the DFT computational cost is really high ($O(N^2)$), where N is the number of computational steps performed by the algorithm, the Fast Fourier Transform (FFT) has been introduced to significantly fasten the procedure (Duhamel et al. 1990). Briefly, the FFT simultaneously computes the odd and even terms of the DFT. As a consequence, the FFT computational cost decreases to $O(Nlog_2N)$. For instance, for $N = 10^9$, if each operation is performed in 1 ns, the same computation is successfully completed by the FFT in 30 s, while the standard DFT needs more than 31 years.

1.2 Phasor Approach

The phasor approach has been originally exploited in spectroscopy by Weber (Weber 1981) and its graphical representation has been introduced some years later (Jameson et al. 1984).

At the state of the art, the phasor approach is widely exploited in biophysics, in particular to analyze fluorescence (Feredouni et al. 2012) and lifetime (Digman et al. 2008) signals.

A description of the method is provided in **Figure 1**.

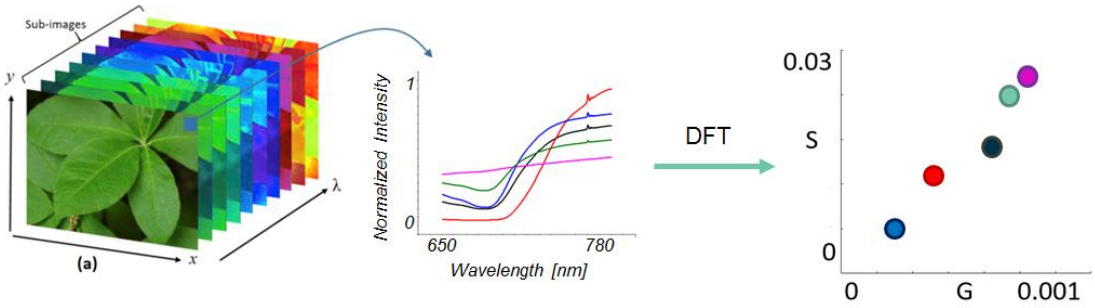

Figure 1. Phasor approach graphical representation. *Each pixel of an hyperspectral image (left) is associated to a single signal (middle), which is projected in a point of the phasor plot (right) with coordinates (G,S) by applying the Discrete Fourier Transform. Signals characterized by different shapes are characterized by different coordinates, while projections of similar functions lie closely in the phasor plot.*

Briefly, the phasor approach takes advantage of the Euler's formula to project a signal into a single point lying in a 2-dimensional space, named phasor plane or phasor plot. The point coordinates represent the real and the imaginary part of the FT. For instance, in an hyperspectral image, each pixel is endowed with a signal f(λ) whose intensity changes in dependence of the acquired emission wavelength. Its real and imaginary coordinates, g(k) and s(k), on the phasor plot are defined as:

$$g(k) = \frac{\int_{-\infty}^{+\infty} f(\lambda)cos\,(2\pi k\lambda)\,d\lambda}{\int_{-\infty}^{+\infty} f(\lambda)d\lambda} \qquad s(k) = \frac{\int_{-\infty}^{+\infty} f(\lambda)sin\,(2\pi k\lambda)\,d\lambda}{\int_{-\infty}^{+\infty} f(\lambda)d\lambda} \qquad (4)$$

In experimental scenarios, **Eq. 4** has to be slightly modified in order to consider that the f(λ) signal intensity is acquired for a discrete number N of equally spaced wavelengths:

$$g\,(k) = \frac{\sum_{\lambda=1}^{N} f(n)cos(\frac{2\pi k\lambda}{N})}{\sum_{\lambda=1}^{N} f(\lambda)} \qquad s\,(k) = \frac{\sum_{\lambda=1}^{N} f(n)sin(\frac{2\pi k\lambda}{N})}{\sum_{\lambda=1}^{N} f(\lambda)} \qquad (5)$$

The g(k) and s(k) coordinates are strictly related to the signal spectral properties and shape (Jameson et al. 1984). Therefore, spectra characterized by similar parameters lie closely within the phasor plot (Scodellaro et al. 2022) and, in particular, spectra with the same shape will have the same phasor plot coordinates. However, during experimental measurements, the acquired spectra are affected by white noise, which impacts the spectral shape and therefore its projection on the phasor space.

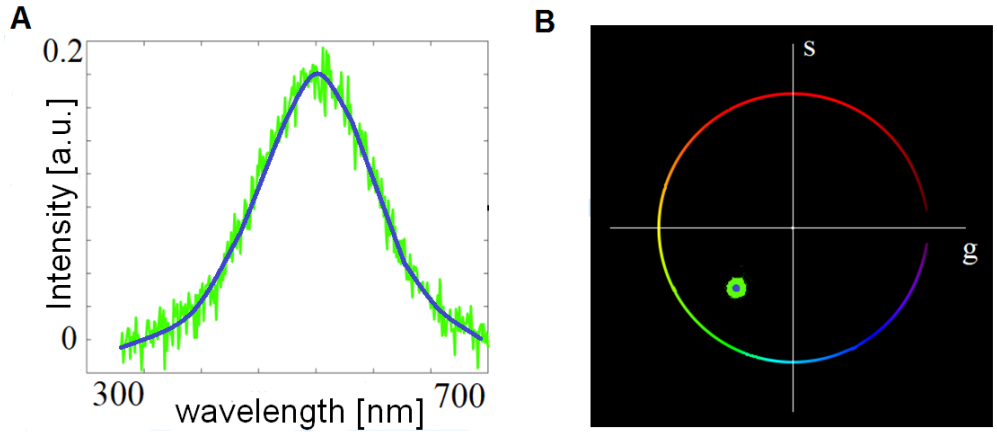

Figure 2. Noise effect on the phasor plot*. **A)** An exemplary function, peaked around 500nm, is represented in blue, while the same function, affected by noise (SNR = 100), is shown in green. **B)** the spectrum not affected by noise is represented by a single point on the phasor plot (in blue), while the green dots are the projection of 10000 spectra characterized by SNR = 100.*

Indeed, as explained in details by Cutrale and colleagues (Cutrale et al. 2017), if (G_1, S_1) is the position on the phasor plot of a spectrum not affected by noise, the projection (G_2, S_2) of the same spectrum affected by a white noise has the probability of 1σ (where sigma represents the distribution standard deviation) (with sigma SCRIVI COS'è) to lie inside a circle of radius r centered in (G_1, S_1) with:

$$r\ \alpha\ \frac{1}{\sqrt{SNR}} \qquad (6)$$

where SNR is the Signal-to-Noise Ratio describing the noise intensity (**Figure 2**). An example is shown in **Figure 2**, where the projection of a single spectrum not affected by noise (**Figure 2A**, in blue), is represented by fixed coordinates on the phasor plot (**Figure 2B**, the blue point). If white noise is applied to the spectrum, the shape on the function is randomly modified (**Figure 2A**, in green). This effect introduces an indetermination affecting the projection position on the phasor plot: this is demonstrated by **Figure 2B**, where the green dots represent the projections of the same spectrum (**Figure 2A**, in blue) when white noise (SNR = 100) is applied on it for 10000 different times. In particular, the green dots are arranged in a circle centered around the projection coordinates of the spectrum not affected by noise (Cutrale et al. 2017). Both i-φ-MaLe and SuperμMAPPS take into account this important property when they analyze experimental spectra (**Section 5.1** and **Section 7.1**).

The phasor approach is characterized by multiple other advantages: since the phasor approach translates entire spectra in single points on the phasor plane, it provides a huge reduction of the dataset dimension, simplifying its handling and interpretation and preserving, at the same time, the majority of the information content. The phasor approach also provides a fast and intuitive representation of the dataset, which is globally analyzed exploiting the dispersion of the data in the phasor space. For instance, differences in the spectral behavior of the sample can be singled-out by segmenting patches of the phasor points distribution directly in the phasor space and by measuring geometrical features of this distribution. More important, the information retrieved graphically is not biased by the initial parameters values and constrains, contrary to the usual non-linear fitting procedure.

1.3 Supervised and Unsupervised Machine Learning

Nowadays, the development of powerful computers and the possibility to access to a large amount of data stored in public repositories have hugely increased the size and the

complexity of statistical problems. Recently, the need of novel techniques concerning data storage led to the birth of a new field called data mining (Jothi et al. 2015). In parallel, there has been an expansion of the data learning field, which is the capability to analyze data by extracting meaningful information, as patterns and trends. These tasks have been automatically accomplished by computers exploiting statistical-based algorithms, leading to the birth of the Machine Learning (ML) field (Mitchell, 1990).

ML is considered a branch of Artificial Intelligence devoted to the development of procedures able to automatically analyze huge amount of data and extract meaningful trends. ML problems can be categorized in two different classes: Supervised and Unsupervised. In Supervised ML (S-ML), the aim is the prediction of an outcome value by exploiting a set of input numbers. In particular, during the first phase, which is called training phase, the parameters characterizing the architecture of the algorithm are set by showing a set of examples, called training dataset. Each entry of the training dataset is composed by the input numbers and the corresponding outcome. The algorithm automatically set its parameters to associate the input number to the corresponding outcome with the best accuracy. In order to validate the method, a set of new data, called test dataset, is exploited during the test phase. These data, composed by the same inputs of the training dataset, are shown to the algorithm for the first time and the approach has to produce an outcome "as it is", without being able to modify its parameters. Then, the algorithm accuracy is evaluated by comparing the prediction provided for each entry of the test dataset with the real outcome, which has been hidden from the method by the user. If few data are available to train the algorithm, k-fold cross-validation is useful to validate our approach. In this case, the data is divided into k subsets. k-1 subsets are exploited as training dataset, and the remaining one is considered the test dataset. Thanks to this approach, every subset is exploited once as test set and k-1 times as part of the training set. The overall accuracy of the ML model is the average value among the k trials. This significantly reduces bias caused by underfitting and variance since a huge number of data is exploited during the test phase. Differently from S-ML, Unsupervised ML (U-ML) does not provide an outcome since its

final goal is finding associations among a set of input data, while it divides data in different groups sharing similar patterns and characteristics.

In particular, the algorithms described in the following sections (**Section 5.1**, **Section 6.1** and **Section 7.1**) are defined as S-ML techniques since their aim is the automated retrieving of the parameters (which correspond to the outcome values) characterizing the analyzed experimental functions, as the Second Harmonic Generation signal for SuperμMAPPS or the reflectance spectrum exploited by i-φ-MaLe.

2. Solar-Induced Fluorescence spectra to evaluate the vegetation stress status: a brief introduction

2.1 The Solar-Induced Fluorescence role in Remote Sensing

When the vegetation is exposed to the sunlight, it absorbs energy, which can be dissipated through different paths: Photochemical Quenching (photosynthetic process), Non-Photochemical Quenching (heat) or fluorescence emission (**Figure 3**).

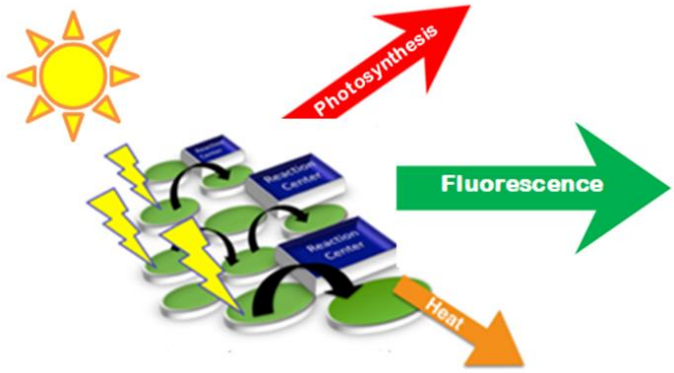

Figure 3. Pathways exploited by vegetation to emit the energy absorbed by sunlight. *In ideal scenarios, only the path related to the photosynthetic process (red arrow) is activated. If the amount of energy coming from the absorbed light is too high, the excess is dissipated through Non-Photochemical Quenching (orange arrow) and fluorescence emission (green arrow).*

The photosynthesis is a crucial process for vegetation since, in experimental scenarios, it allows to avoid damage to the photosynthetic apparatus caused by an excessive amount of sunlight reaching the canopy (Pinto et al. 2020). Instead, fluorescence (F) is an important marker that provides a quantitative description and characterization of potential stresses affecting the vegetation (Porcar-Castell et al. 2014, Hendrickson et al. 2004, Baker 2008, Mohammed et al. 2019), in particular if related to the analysis of some biophysical parameters (Malenovski et al. 2009), as the Leaf Area Index (LAI) or the Chlorophyll Content (Cab).

As already reported in literature (Cogliati et al. 2019), F can be detected by exploiting spectrometers at different spatial scales, from ground to satellite distances (Mohammed et al. 2019) from the canopy, as shown in **Figure 4**. In particular, the possibility to retrieve F from the hyperspectral signals acquired from satellite, target of the ESA FLEX mission 2025 (De

Grave et al. 2020), paves the way towards the supply of global maps of vegetation fluorescence, which are crucial in order to better understand the global photosynthetic activity and assist agricultural management.

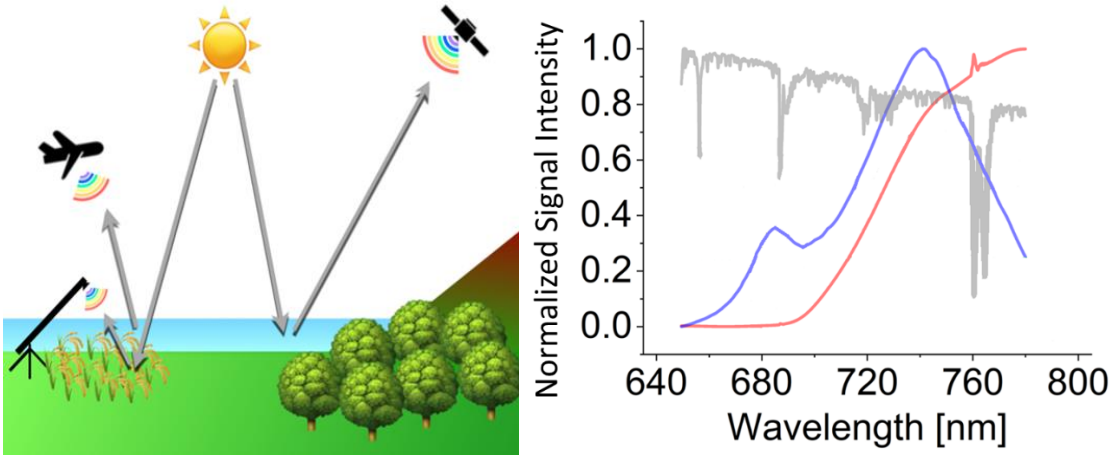

Figure 4. Exploited spatial scales for fluorescence detection in the Remote Sensing field and fluorescence-acquisition related spectra representation. *The excess of absorbed energy is dissipated by vegetation via fluorescence emission (represented by the blue line). Fluorescence can be detected by spectrometers which are positioned at top of canopy level or mounted on tower or airplanes. Recently, the detection of fluorescence signals from satellites is the main target of important international projects, as the ESA FLEX mission. In particular, the relation between* L^{up} *(grey line),* R_{app} *(red line) and F spectra (blue line) is exploited.*

Theoretically, the fluorescence signal F is related to other quantities:

$$L^{up}(\lambda) = R(\lambda)L^{dw}(\lambda) + F(\lambda) \qquad (7)$$

where L^{up} (grey line in **Figure 4**) is the whole radiance measured over the canopy of the vegetation under investigation, while R represents the canopy reflectance contribution. This quantity is multiplied to the reflectance contribution of elements other than the tree canopy (L^{dw}). However, experimentally, through spectrometers it is possible to retrieve both L^{up} and

L^{dw}. As a consequence, the apparent reflectance R_{app} (red line in **Figure 4**) can be obtained by rearranging **Eq. 7**:

$$R_{app} = \frac{L^{up}(\lambda)}{L^{dw}(\lambda)} = R(\lambda) + \frac{F(\lambda)}{L^{dw}(\lambda)} \qquad (8)$$

It is trivial to notice that the apparent reflectance is related to fluorescence (blue line in **Figure 4**). However, fluorescence detection by remote sensing techniques is really challenging since it represents only a small portion (less than 1%) of the total signal acquired by the sensor in the range [650nm 780nm]. Moreover, compared to the R_{app} spectrum, F is originated only by vegetation and it is more related to plant physiology (Verrelst, et al., 2015). Indeed, F intensity strictly depends on some biophysical parameters as the Fluorescence Quantum Yield (Fqe) and the vegetation Chlorophyll Content (Cab), which quantify the efficiency of fluorescence emission by the leaves and the amount of chlorophyll characterizing the canopy leaves, respectively, and whose detection is crucial to investigate vegetation stress status (**Figure 3**). Due to their reciprocal dependence, both R_{app} and F spectra shapes are trivially dependent on Fqe and Cab. R_{app} and F also depend on other biophysical parameters, as the leaf area and composition and the fraction of solar light absorbed by the chlorophyll pigments. This has been demonstrated by recent studies (Van der Tol et al. 2009, Van der Tol et al. 2014). Moreover, the Absorbed Photosynthetically Active Radiation (APAR), which is the solar radiation absorbed by leaves for the photosynthesis activity, is dependent on the R_{app} signal since it can be computed by integrating L^{up} and L^{dw}:

$$APAR = |f_s| \int_{400nm}^{700nm} L^{dw}(\lambda)[1 - \rho(\lambda) - \tau(\lambda)]d\lambda + \int_{400nm}^{700nm} L^{up}(\lambda)[1 - \rho(\lambda) - \tau(\lambda)]d\lambda \qquad (9)$$

Eq. 9 also demonstrates the APAR dependence on the leaf reflectance (ρ) and transmittance (τ). Notably, since the parameter f_s takes into account the leaf orientation compared to the sunlight, the leaf area index (LAI) parameter is able to affect the R_{app} spectrum.

Another important parameter is the Solar Zenith Angle (SZA). Indeed, as shown in **Figure 5**, the entire R_{app} signal decreases for high SZA values since the illumination provided by the sun is less intense if the light is not perpendicular to the canopy.

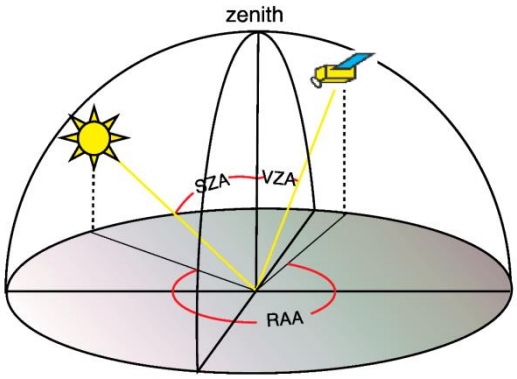

Figure 5. Comparison between Solar Zenith Angle (SZA) and Sensor Zenith Angle (VZA). *For higher SZA values, the R_{app} intensity decreases while the angle between the direction of the light coming from the sun and the perpendicular to the vegetation canopy increases.*

Moreover, the Fluorescence signal F is strongly influenced by the canopy itself: indeed, the total fluorescence produced by the photosystem, which is called F_{RC}, is re-absorbed by the leaf causing a decrease of the signal F emitted by the vegetation (10 to 100 times intensity reduction). Unfortunately, this absorption is dependent on the considered spectral wavelength. In particular, the peak around 687 nm is strongly reduced, leading to a difference in the F_{RC} and F spectra shape.

In order to quantitatively explain all these complex dependencies, different Radiative Transfer (RT) models have been recently designed. The Soil Canopy Observation of Photosynthesis and Energy fluxes (SCOPE) 2D model (Van del Tol et al. 2009) simulates the vegetated land surface fluxes, connects them to the satellite observations characterizing the visible, infrared and thermal spectral domains by applying physical models and

quantifies the micro-climate in the canopy. In a more experimental perspective, among the numerous input values of the SCOPE model, the LAI, Cab, SZA and Fqe parameters can be tuned, while the RT architecture provides the APAR and the corresponding R_{app}, F and F_{RC} spectra. At the state of the art, 3D-based RT models, as DART (Malenovski et al. 2009, Gastellu-Etchegorry et al. 2017), have been designed, even if they are not widely exploited in the experimental research field due to their extremely higher complexity compared to SCOPE. Moreover, theoretical models devoted to the quantification of the atmospheric transmission have been developed, as MODTRAN (MODerate resolution atmospheric TRANsmission, Berk et al. 2014). These different approaches can be synergistically combined to accurately describe the whole radiative transmission phenomenon acting on vegetation canopies, as demonstrated by Cogliati and colleagues, who exploited the results of the simulations obtained by MODTRAN to improve the SCOPE outputs (Cogliati et al. 2019).

This really complex scenario is even more complicated if tower or satellite-level measurements are considered. Indeed, if the distance between the vegetation canopy and the spectrometer increases (from 100 m and beyond), the R_{app} acquired by the instruments is affected by a huge modification in its spectral shape due to reabsorption effects caused by the telluric atmospheric oxygen. In particular, O_2 bands around 687 nm and 760 nm, which are the most informative part of the spectrum concerning F (Mohammed et al. 2019), are the most affected spectral regions. Nowadays, this issue is one of the most crucial challenges regarding the remote sensing field since the majority of the information concerning F is not available from satellite-level experimental acquisitions.

2.2 State of the art of the Fluorescence spectrum deconvolution

In the remote sensing field, the dependence among R_{app}, F and the other biophysical parameters described in the previous **Section 2.1** has been widely exploited to retrieve the F spectrum from the measured R_{app} (Mohammed et al. 2019). Unfortunately, the

disentanglement of the F spectrum from the canopy reflected spectral radiance (Cendrero-Mateo et al. 2019, Meroni et al. 2009) and from the atmospheric influence (Sabater et al. 2018, Aasen et al. 2019) is a challenging task of the Remote Sensing field (Zhao et al. 2018).

At the state of the art, the methods devoted to the retrieval of the F spectrum from the spectral radiance reflected by the canopy (Chang et al. 2020, Cogliati et al. 2018, Grossmann et al. 2018, Frankenberg et al. 2018. Siegmann et al. 2019) are based on inversion models, which require high-spectral resolution observations at 1) the telluric O_2 absorption bands (687 and 760nm), not suitable to analyze satellite-level measurements since they are very sensitive to the atmospheric signal reabsorption, or at 2) the Solar Fraunhofer Lines, which are minimally influenced by atmospheric reabsorption phenomena affecting F_{RC} but contain less information concerning F (Mohammed et al. 2019).

Spectral Fitting Methods SFM (Meroni et al. 2010, Mazzoni et al. 2010, Cogliati et al. 2015b) and SpecFit (Cogliati et al. 2019) have been demonstrated the most robust inversion-based methods for the fluorescence signal retrieving. Briefly, these approaches are based on a fitting procedure applied to the O_2 bands of the analyzed R_{app} function (around 687 nm and 760 nm). In particular, the corresponding F is retrieved by inversely exploiting the SCOPE model, which computes R_{app} from the F spectrum. SFM retrieves F at 687 nm and 760 nm, the two oxygen bands, while SpecFit can be considered its extension since it extracts the entire F spectrum in the range [650nm 780nm]. Unfortunately, these methods are useless for satellite-based measurements, as the acquired R_{app} spectrum is modified by the atmospheric oxygen reabsorption. More recently, Artificial Intelligence-based techniques have been discovered as powerful tools to analyze data related to hyperspectral imaging space missions, as ESA FLEX, NASA SBG (Cawse-Nicholson et al. 2021) or ASI PRISMA (Verrelst et al. 2021). However, the coupling between novel machine learning techniques and state of the art Radiative Transfer (RT) models, able to theoretically derive the F spectrum by considering both the vegetation-related biophysical parameters (as LAI, Cab,

the leaves shapes) and the expected environmental conditions (as the Solar Zenith Angle), has never been addressed for fluorescence studies, although it is expected to provide new insights for detecting vegetation status.

Within this challenging context, a new approach based on a fast optimized multi-parameter retrieval algorithm, named i-φ-MaLe (hybrid Phasor-Machine Learning based approach), has been modeled. i-φ-MaLe is specifically aimed to retrieve the F spectrum, the F_{RC} spectrum, F quantum efficiency (Fqe) and canopy biophysical parameters consistently. The proposed algorithm consists of a novel hybrid Phasor-Machine Learning based approach, which is developed here and exploited for the first time for quantitative retrievals in the context of Remote Sensing studies. In particular, i-*φ*-MaLe is able to compare the shapes of the apparent reflectance spectra (R_{app}), acquired by state of the art spectrometers, with a wide set of Radiative Transfer (RT) simulations obtained by coupling SCOPE and MODTRAN architectures (Cogliati et al. 2019) to detect the function which best matches the experimental data. Then, the method associates the biophysical and fluorescence parameters characterizing the simulated function to the analyzed experimental spectrum. Thanks to this approach, a complete characterization of the vegetation has been performed for the first time by exploiting a single algorithm able to simultaneously provide both F spectra and many biophysical parameters.

3. Challenges and opportunities of AI-based tools in histopathology

Biopsies play a key-role during the clinical routine as they are a crucial tool to provide accurate diagnosis. In the standard clinical procedures, tissue samples are excised from patients and directly observed under the microscope by pathologists after being fixed, cut and stained with dyes that label, more or less specifically, different biological structures. Hematoxylin and Eosin (H&E) is one of the most exploited staining in the clinical practice (Fischer et al. 2008, Cardiff et al. 2014). In particular Eosin, characterized by a magenta color if observed under transmitted light, stains proteins, while Hematoxylin, which appears deep blue, highlights nucleic acids. Moreover collagen, a protein whose quantification is crucial during fibrosis and tumor diagnosis, appears red after the Picrosirius Red staining (Lattouf et al. 2014).

During the biopsies analysis, the pathologist provides a semi-quantitative characterization of the stained biological structures by describing their morphology, amount and relative positions (Gurcan et al. 2009).

Recently, the introduction of whole-slide imaging scanners (Pantanowitz et al. 2011) and the consequent spread of public repositories (where a huge amount of high-resolution images of stained human tissue sections are collected - Lopez et al. 2012, Tomczak et al. 2015, Wang et al. 2016, Clark et al. 2013) paved the way to a wide exploitation of artificial intelligence-based protocols in healthcare.

3.1 Image acquisition and storage

At the state of the art, the images commonly exploited in histopathology are acquired at high resolution and magnification (pixel dimension ~0.4 μm) by means of whole-slide scanners. As a consequence, these images require a lot of memory space to be stored (in the order of GBs for each sample) and, consequently, loaded. In order to speed up their loading process, the images are saved in pyramidal format, shown in **Figure 6**.

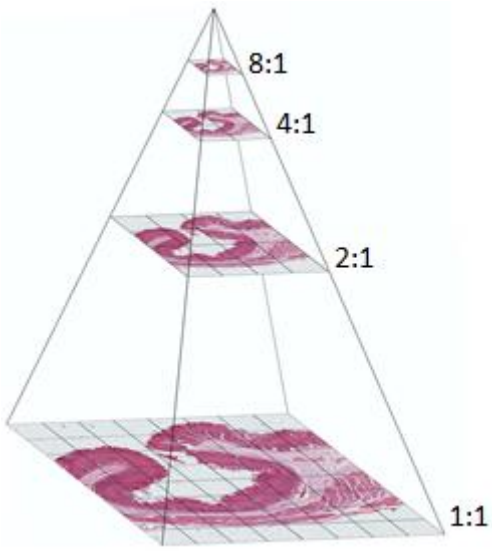

Figure 6. Architecture of a pyramidal image. *In each layer of the pyramid the whole slide is saved resampled at decreasing magnification and therefore resolution.*

In this regard, a pyramidal image is characterized by many different layers, where the entire sample is saved with a decreasing magnification and therefore resolution. In particular, each layer is composed by square image tiles of the same dimensions (e.g. 1024x1024 pixels) and the bottom layer is a 1-to-1 pixel mapping of the original acquired data. The upper layers are obtained by downsampling the original image: for each level the pixel size is doubled. Thanks to this approach, global analysis of the sample can be performed by exploiting a level characterized by a lower resolution, with a consistent reduction of the time needed to load the image. When a full-resolution reconstruction of the sample is needed, lower layers of the pyramidal format are selected and employed.

In this scenario, pyramidal images are even more useful since whole slide scanners store the saved image in RGB format, which need more memory than traditional grayscale images. Indeed, in this format, each image of dimension [M x N] is represented by a 3-level hypermatrix [M x N x 3], where the three levels correspond to the red, green and blue image channels (**Figure 7**).

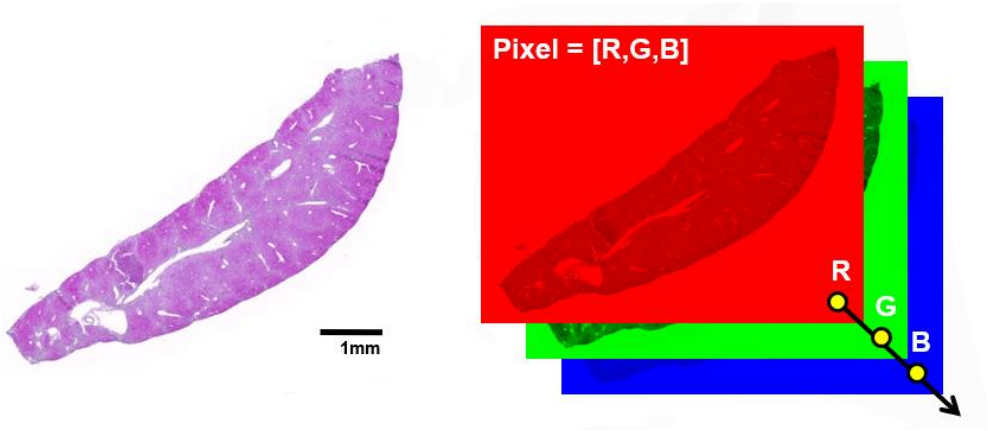

Figure 7. H&E-stained biopsy represented in the RGB format*. Each pixel is associated to three numbers, which correspond to the intensity of the red, green and blue contribute characterizing its color.*

As a consequence, in the RGB color space, each pixel is associated to three different numbers (one for each color channel) in the interval [0 2^b], where b is the number of bits exploited to store a single pixel, which represents the intensity R, G and B of the red, green or blue contribute of the pixel color, respectively. These values can be represented by a vector of three values [R, G, B]. Thanks to this representation, a 8-bit RGB image, which is the most common format exploited by state-of-the-art whole-slide scanners, is able to assign to each pixel one of more than 16 millions colors, represented by a different [R, G, B] vector.

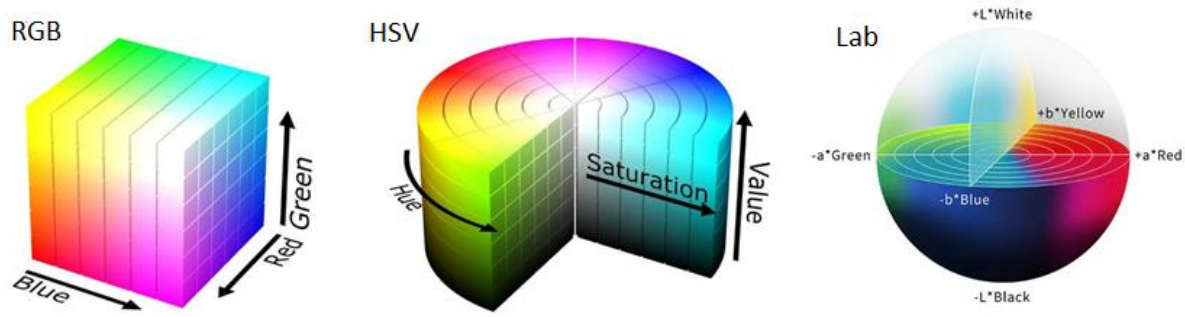

Figure 8. RGB, HSV and laβ color spaces*. Each pixel represented in the RGB format can be converted in the HSV or laβ color space.*

As reported in the literature (Hanumantharaju et al. 2011, Saravanan et al. 2016), each pixel expressed in the RGB format can be mapped to other 3D representations, able to highlight different colour properties. One of the most exploited is the HSV (Hue, Saturation and

Value) space, shown in **Figure 8**. In particular, the H, S and V parameters characterizing each pixel can be expressed in dependence of the [R, G, B] format as (Chernov et al. 2015):

$$H = \{\frac{G-B}{max_{RGB}-min_{RGB}} \quad if\ R = max_{RGB}$$

$$H = \frac{2+(B-R)}{max_{RGB}-min_{RGB}} \quad if\ G = max_{RGB} \qquad (10)$$

$$H = \frac{4+(R-G)}{max_{RGB}-min_{RGB}} \quad if\ B = max_{RGB}$$

$$S = \frac{max_{RGB}-min_{RGB}}{max_{RGB}} \qquad (11)$$

$$V = max_{RGB} \qquad (12)$$

Where max_{RGB} and min_{RGB} are the highest and the lowest values among the R, G and B variables. The HSV space emphasizes the discrepancies in the color chromaticity by exploiting the Hue, which is independent of white or black contributes, described only by the Saturation and the Value variables.

3.2 Staining normalization algorithms role in Digital Pathology

When the image of a biopsy is acquired and stored by whole-slide scanners, it can be analyzed by means of AI-based methods to provide both the automatic extraction of relevant features characterizing the diseases and the quantitative evaluation of biological structures (Bi et al. 2019), overcoming the drawbacks of a subjective visual inspection to infer diagnosis. Despite Machine Learning-based approaches are able to effectively assist pathologists during their daily routine by reducing their workload and providing quantitative data to support the diagnosis, these techniques are often based on the analysis of features strictly related to the color of each stained biological structure (Bug et al. 2015, Kleczek et al. 2020). In addition, even if some ML protocols seem to be not affected by the image color content, as they are based on shape and textural features, the segmentation exploited to compute the morphology-based properties strictly depends on the proper choice of a color threshold (Bakheet et al. 2021, Nemat et al. 2018). Moreover, the color content also affects

both textural properties, which are extracted from matrices describing the relative position among pixels characterized by different color intensities (Haralick et al. 1973, Mohaniah et al. 2013), and Neural Networks-based protocols (Vu et al. 2019, Coudray et al. 2018, Van Rijthoven et al. 2021, Anghel et al. 2019), which exploit the image colors as input data.

Unfortunately, staining is an experimental procedure which suffers of inhomogeneity in the color highlighting the same biological structures, both in the same and in different tissue sections (Van Seijen et al. 2019). This is due both to the performed staining procedure and to the image acquisition by means of different scanners. Indeed, the colors characterizing the digitized images acquired by the whole-slide scanners are affected by the illumination conditions of the sample and the color response of the digital camera. These drawbacks interfere with the performance of Artificial Intelligence-based analysis of tissue sections prepared and acquired in different laboratories. A straightforward example of this crucial problem is shown in **Figure 9**. Here, H&E-stained patches of the same tissue (liver), acquired in different laboratories, have been analyzed by applying the same Machine Learning procedure, which is a simple thresholding method. In particular, the algorithm selects the pixels characterized by a blue contribute intensity lower than 0.65 and associates them to the nuclei structures, highlighted in green in **Figure 9B** and **Figure 9D**.

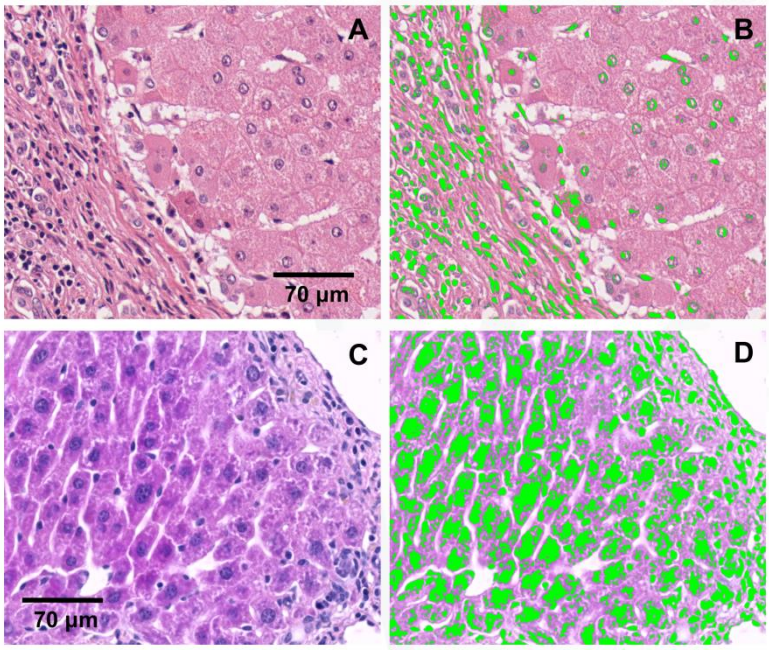

Figure 9. Results of a color-based nuclei segmentation procedure applied to two H&E-stained images acquired in different laboratories. *Panel **A** shows a patch of a human liver sample, H&E stained and acquired at the laboratories of the University of Leeds (available in their public repository, ULVPR https://www.virtualpathology.leeds.ac.uk/). Panel **C** reports a patch of an H&E-stained murine liver sample acquired at the DKFZ laboratories of Heidelberg. The nuclei of each patch have been segmented by applying the same Machine Learning algorithm, which selects pixels characterized by a blue contribute intensity lower than 0.65. The selected pixels are highlighted in green in Panels **B** and **D**. Stain normalization methods have not been exploited.*

The patches, clearly characterized by different colors, highlight a different performance of the algorithm: the image acquired at the laboratories of the University of Leeds (available in their public repository – ULVPR) (**Figure 9A**) shows a much higher accuracy in nuclei segmentation (**Figure 9B**) with respect to the sample acquired at the DKFZ laboratories of Heidelberg (**Figure 9D**), where the algorithm selects many pixels belonging to the cells cytoplasm.

This example clearly highlights the need of a biunivocal correspondence between the considered biological structures and their color in the image to successfully apply Artificial Intelligence protocols. As a consequence, different staining normalization methods have been developed to reach this target. In particular, these protocols uniform the color expressed by each staining in images of tissue sections acquired in different experimental conditions (Reinhard et al. 2001, Macenko et al. 2009, Khan et al. 2014, Vahadane et al. 2016, Shaban et al. 2019). A complete description of advantages and limitations of already

developed stain normalization methods has been recently provided by Lakshmanan and colleagues (Lakshmanan et al. 2019). Among these approaches, the most exploited have been developed by Reinhard and colleagues (Reinhard et al. 2001) and Macenko and colleagues (Macenko et al. 2009). Reinhard's method is based on color transfer between a reference image and a target image, which are transformed to lαβ space (**Figure 8**) during a pre-processing procedure. In particular, the color normalization is performed by transferring the mean and variance values of every image channel from the reference to the analyzed image (Reinhard et al. 2001). Macenko's pipeline finds the specific stain vectors describing the color expression of the analyzed image in the RGB space and maps each pixel color from the retrieved domain to the domain of a fixed set of stain vectors elected by the user (Macenko et al. 2009). Unfortunately, the Reinhard's approach is not robust concerning the stains separation, while Macenko's procedure provides satisfying results only for images characterized by a less intense staining. In order to mitigate these issues, I developed a new geometrical method, named Φ-Norm (Phasor-based color normalization), which exploits the DFT geometrical properties in order to perform the stain normalization task. Further theoretical details about Φ-Norm, its validation and applications in experimental scenarios are reported in **Section 6**.

4. Label-free Digital Pathology

Despite playing a key-role in assisting pathologists during the diagnosis, staining is an expensive and time-consuming procedure (Rivenson et al. 2020). To overcome this issue, recently, many attempts have been made to predict the staining of tissue sections through Neural Networks (Rivenson et al. 2019). An interesting alternative to staining protocols consists in taking advantage of label-free microscopy techniques (Fereidouni et al. 2012, Radaelli et al. 2017, Scodellaro et al. 2019, Mazumder et al. 2019), which exploit the autofluorescence or scattering signal generated by some endogenous proteins. Label-free techniques can overcome many drawbacks characterizing the standard fluorescence-based imaging procedures: the extensive tissue manipulation that can lead to artifacts, the lack of 3D sectioning and the delay in obtaining the diagnosis. Thanks to these approaches, some biological structures which are highlighted by staining procedures, as collagen, are easily visualized without the need of any extrinsic dye.

In this regard, collagen is the most abundant protein in tissues, where it provides structural support and protection (Buehler et al. 2006, Cicchi et al. 2013). Its structure is reported in **Figure 10**.

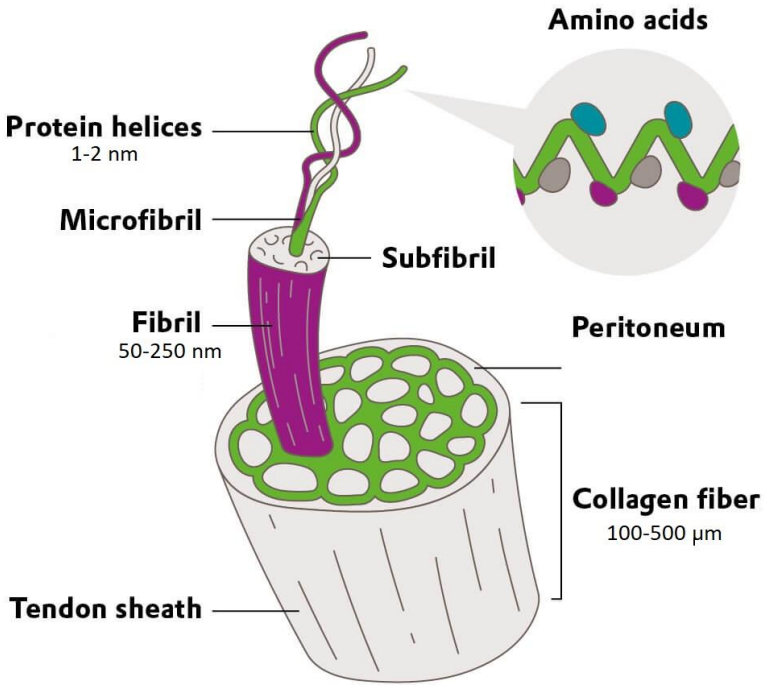

***Figure 10. Collagen structure**. Each collagen fiber is composed by subunits, named fibrils, which are the result of the bond among many protein helices. Each helix is characterized by three α-chain. The original image is available at: https://elements.evonik.com/research-and-Innovation/A-versatile-helper/*

Collagen synthesis begins in the fibroblasts, where peptide subunits of hundreds of residues are assembled on the ribosome and linked to form α-chains. Three α-chains bond together to create collagen fibrils (Bella et al. 1994), which are the subunits of the collagen fiber.

Collagen micro-architecture, both at fibrils and fibers resolution level, changes during pathological processes. As a consequence, it has been successfully exploited as an early diagnostic marker both in experimental and clinical medicine (Raja et al. 2015, Liao et al. 2013, Wang et al. 2015). In particular, collagen can be successfully label-free imaged and characterized thanks to the Second Harmonic Generation (SHG) signal. Moreover, the polarization dependence of the SHG response (P-SHG) has been demonstrated to be useful in order to extract features, as the collagen fibrils orientation and disorder, related to the tumor growth stage (Radaelli et al. 2017, Scodellaro et al. 2019) and to effectively describe the collagen supramolecular structure.

Recently, the P-SHG signal has been seriously considered also for clinical applications, thanks to the presence of novel public repositories, as PSHG-TISS (Hristu et al. 2022, https://doi.org/10.17605/OSF.IO/UDTQP), where hundreds of P-SHG images belonging to different kind of human tissues have been provided and classified. Moreover, the presence of public repositories could pave the way to the application of supervised Machine Learning algorithms able to analyze the features extracted from the P-SHG signals.

4.1 Second Harmonic Generation Microscopy

SHG is a coherent non-linear optical process induced in birefringent crystals or in biological structures characterized by a non-centrosymmetric symmetry , as collagen. Two incident photons at their fundamental frequency interact with the non-centrosymmetric medium and are directly converted during a quantum event into a single photon characterized by the same total energy and half wavelength, without the absorption or emission processes (**Figure 11**). SHG is characterized by several advantageous properties which make this microscopy technique a unique tool for imaging tissues in situ (Campagnola et al. 2012). At first, SHG is a scattering phenomenon, produced from endogenous proteins, as collagen and myosin. As a consequence, photobleaching and phototoxicity are avoided compared to standard fluorescence-based imaging methods. Thanks to the near infrared range of the exploited laser wavelengths, SHG guarantees deep tissue imaging (Campagnola et al. 2002). Moreover, SHG signal is strongly dependent on the structural characteristics of the tissue, as briefly explained in the following (Campagnola et al. 2012 for a more extensive discussion).

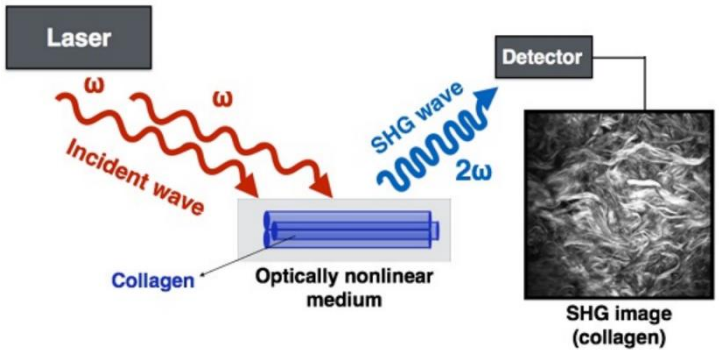

Figure 11. Scheme of the SHG process. *Two photons of frequency ω interact with collagen, which is an optically nonlinear medium. The scattering process produces a single photon of frequency 2ω, which is the SHG signal acquired by a detector.*

The polarization P of a non-centrosymmetric structure is non-linear and expressed by:

$$P = \varepsilon_0\left(\chi_{ij}^{(1)}E + \chi_{ijk}^{(2)}E^{(2)} + \cdots + \chi_{ijkl\ldots n+1}^{(n)}E^{(n)}\right) = \varepsilon_0\chi_1 E + P_{nl} \qquad (13)$$

Where P_{nl} is the polarization non-linear term and E is the electric field vector. Due to the second order nature of SHG, only the non-linear susceptibility tensor χ_{ijk} of dimensions 3x3x3 can be considered non-zero in the P_{nl} term. χ_{ijk} is strictly dependent on the density of molecules N_s and their mean hyperpolarizability orientation β:

$$\chi_{ijk}^{(2)} = N_s < \beta > \qquad (14)$$

Eq. 14 demonstrates that the susceptibility term is non-zero in presence of a non-centrosymmetric structure, as collagen (described by a hexagonal or cylindric symmetry, as shown in **Figure 10**). As a consequence, the second-order induced polarization term responsible for SHG can be easily derived (Campagnola et al. 2002):

$$P_{SHG} = 2\varepsilon_0 \chi_{ijk}^{(2)} \qquad (15)$$

where ε_0 is the susceptibility. Under semiclassical approach, starting from Maxwell's equations, the equation for field propagation $A_{2\omega}$ is:

$$\frac{\partial A_{2\omega}}{\partial z} = -j \frac{\mu_0 \omega_0 c}{2n} P e^{-j\Delta k z} \qquad (16)$$

where Δk is the phase matching term. The measured intensity of the SHG process is given by:

$$I_{SHG} = |A_{2\omega}|^2 = \alpha_{2\omega}^2 |\chi_{ijk}^{(2)}|^2 I_\omega^2 L^2 sinc^2(\frac{\Delta k L}{2}) \qquad (17)$$

Where $\alpha=\omega_0/nc$ (where c is the speed of light, ω_0 the wavelength frequency) is constant and L is the sample length. **Eq. 17** demonstrates that the SHG signal intensity depends on the fibrils orientation and it is maximum if the fibers are aligned on the plane perpendicular to the laser propagation direction (Lacomb et al. 2008). The dependence on the polarization of the I_{SHG} signal suggests that it can be exploited for polarization-dependent measurements (Polarized Second Harmonic Generation Microscopy, P-SHG). The P-SHG mathematical description has been extensively provided in literature (Campagnola et al. 2002, Plotnikov et al. 2006, Radaelli et al. 2017).

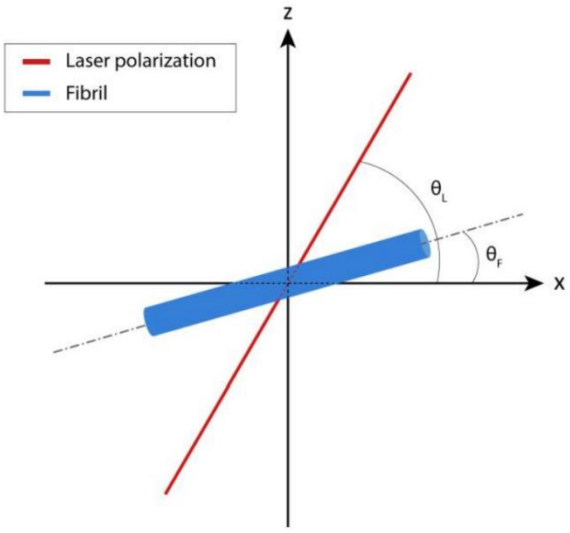

Figure 12. Scheme for the P-SHG mathematical description. *The collagen fibril lies in the xz plane. Its orientation is described by the θ_F angle. The laser beam propagation direction is along the y-axis, while its polarization angle is described by θ_L.*

Briefly, as showed in **Figure 10**, the collagen fibrils structure can be described by a hexagonal or cylindrical symmetry. Therefore, Kleinman conditions (Wagniere 1986) can be applied on the susceptibility term:

$$\chi_{ijk} = \chi_{ikj} \quad (18)$$

$$\chi_{xxz} = \chi_{xzx} = \chi_{yzy} = \chi_{yyz} = \chi_{zyy} = \chi_{zxx} \quad (19)$$

If the collagen fibril lies in the xz plane and the propagation direction of the laser beam is along the y-direction, as reported in **Figure 12**:

$$\chi_{zxx} = \chi_{zyy} \quad (20)$$

$$\chi_{xxz} = \chi_{xzx} = \chi_{yzy} = \chi_{yyz} \quad (21)$$

$$\chi_{xzy} = \chi_{yzx} = -\chi_{yxz} = -\chi_{yzx} \quad (22)$$

As a consequence, in this scenario, the non-linear polarization of an electric field E linearly polarized with an angle θ_L-θ_F (where θ_F and θ_L are the angles between the x-axis and the fibril and the laser polarization orientation, respectively) is described in dependence of only two components χ_{zxx} and χ_{zzz} of the susceptibility tensor:

$$P_{nl} = [2\chi_{zxx}Esin(\theta_L - \theta_F)Ecos(\theta_L - \theta_F)]x + \chi_{zxx}[Esin(\theta_L - \theta_F)]^2 + \chi_{zzz}[Ecos(\theta_L - \theta_F)]^2 z \quad (23)$$

Since the total P-SHG intensity signal is proportional to $|P_{nl}|^2$ (Roth et al. 1978, Tiaho et al. 2007):

$$I_{PSHG} = k\left\{sin[2(\theta_L - \theta_F)] + \left[sin^2(\theta_L - \theta_F) + \frac{\chi_{zzz}}{\chi_{zxx}}cos^2(\theta_L - \theta_F)\right]\right\} \quad (24)$$

where the ratio $\gamma = \frac{\chi_{zzz}}{\chi_{zxx}}$, in case of collagen, is a parameter able to quantitatively describe the overall disorder of the fibrils and the scale factor k includes the absolute intensity of the P-SHG signal and setup parameters (Campagnola et al. 2002), such as laser power, optical transmission ratios for excitation signal light path and for the detector sensitivity. **Eq. 24** is valid only for collagen fibers lying parallel to the surface of the tissue (i.e. the collagen fibrils and the incident electric field lie within the focal plane), otherwise the Kleinman symmetry conditions are not verified.

Recently, the I_{PSHG} signal has been described by a different model, which considers non-null the component χ_{zxz} (Amat-Roldan et al. 2010, Psilodimitrakopoulos et al. 2016):

$$I_{PSHG} = k\left\{sin[2(\theta_L - \theta_F)] + \left[\frac{\chi_{zxx}}{\chi_{zxz}} sin^2(\theta_L - \theta_F) + \frac{\chi_{zzz}}{\chi_{zxz}} cos^2(\theta_L - \theta_F)\right]\right\} \qquad (25)$$

In this scenario, the collagen fibril disorder is described by two different parameters, $\gamma_1 = \frac{\chi_{zxx}}{\chi_{zxz}}$ and $\gamma_2 = \frac{\chi_{zzz}}{\chi_{zxz}}$.

From an experimental point of view, in the P-SHG microscopy, the I_{PSHG} signal characterizing the collagen fibrils present in the tissue is acquired by means of an optical setup characterized by a pulsed laser, commonly exploited for two-photon excitation microscopy (Oheim et al. 2006), whose polarization is modified along the optical path by a halfwave plate. As a consequence, each pixel of the image stack is associated to a signal, obtained by acquiring the P-SHG intensity for different polarization angles θ_L in the range [0 π]. Then, the P-SHG signals can be analyzed by exploiting **Eq. 24** or **Eq. 25** during a standard fitting procedure able to pixelwise retrieve θ_F and γ or θ_F, γ_1 and γ_2, respectively (Hristu et al. 2022). Unfortunately, fitting methods are quite accurate but time-consuming, limiting the possibility to analyze wide fields of view. During the last years, fit-free methods have been developed to reduce the computational time, preserving the reliability of the retrieved parameters. Among these approaches, µMAPPS (Radaelli et al. 2017) retrieves θ_F by applying the Discrete Fourier Transform to the P-SHG signal and analyzing its projection position on the phasor plot. Then, it applies a second Discrete Fourier Transform to the analyzed spectrum by fixing θ_F and estimates γ. This method is characterized by really low computational costs: as a consequence, entire whole-slide samples have been analyzed by exploiting µMAPPS (Scodellaro et al. 2019).

However, the experimental acquisition of P-SHG signals is still quite slow compared to other imaging techniques since at least 10-15 images must be acquired by finely sampling the polarization angle θ_L to perform an accurate analysis (Hristu et al. 2016, Radaelli et al. 2017).

In this scenario, we conceived a novel method, named SuperμMAPPS (Supervised Microscopic Multiparametric Analysis by Phasor projection of Polarization-dependent SHG), to hugely reduce the P-SHG image stack acquisition time. Thanks to the coupling between the standard phasor approach and a novel machine learning approach acting on the phasor space, we limited the number of experimental points of the P-SHG curves to 6, proving that they are sufficient to provide reliable quantifications (further information in **Section 7.2**).

5. i-Φ-MaLe application in the Remote Sensing field

In this chapter, the architecture of the i-Φ-MaLe algorithm and its application to experimental acquired R_{app} spectra are reported. At first, the novel proposed approach is described and its characterization in terms of retrieval accuracy is reported by exploiting simulated R_{app} spectra obtained from the MODTRAN5+SCOPE radiative transfer model (further information in **Section 2.1**). In particular, i-Φ-MaLe performances have been quantified for different algorithm configurations (i.e. by modifying its hyperparameters, as the width of the exploited spectral windows or the superposition between consecutive windows) and experimental scenarios (i.e. by varying the amount of noise affecting the analyzed spectra) in terms of Relative Root Mean Square Error (RRMSE).

Then, the algorithm has been exploited to analyze real data acquired at top of canopy level (~ 1-3 m between the vegetation and the spectrometer) and assess the seasonal and diurnal trends of each variable characterizing two crops (forage and Alfalfa). The obtained results have been compared with the output of other published algorithms devoted to the retrieval of F spectra in order to investigate their consistency. Finally, seasonal and daily sets of R_{app} spectra acquired at tower-level (~ 100 m between the vegetation and the spectrometer) over a Downy oak forest in France have been analyzed in order to demonstrate the capability of i-Φ-MaLe in providing consistent results also in presence of atmospheric oxygen reabsorption effects which modify the R_{app} shape around 687 nm and 760 nm.

5.1 i-Φ-MaLe algorithm

Method overview

i-*φ*-MaLe (Scodellaro et al. 2022) is a novel multi-parameter retrieval method, based on a modified version of the standard spectral phasor approach (Fereidouini et al., 2012, **Section 1**). It has been conceived and developed in the frame of the Remote Sensing field to consistently estimate fluorescence spectra at top-of-canopy (F), fluorescence spectra at photosystem level (F_{RC}) and 4 different biophysical parameters (Leaf Area Index – LAI,

Chlorophyll Content – Cab, Fluorescence Quantum Yield – Fqe, Absorbed Photosynthetically Active Radiation – APAR) from canopy R_{app} spectra.

Compared to the standard phasor approach, commonly exploited to analyze fluorescence lifetime (Digman et al. 2008, Malacrida et al. 2021) and spectral (Fereidouni et al. 2012, Cutrale et al. 2017) images, i-φ-MaLe introduces three crucial novelties:

1. For the first time, the Fourier Transform is not computed on the entire spectrum, but on consecutive fixed spectral windows. As a consequence, each analyzed R_{app} spectrum is converted in a single point lying on a different phasor space for each spectral interval;
2. A supervised ML algorithm associates to each point in the phasor space the fluorescence spectra and the biophysical parameters characterizing the projected R_{app} spectrum, simulated by exploiting a Radiative Transfer model.
3. The fluorescence spectra and the biophysical variables of experimental R_{app} are obtained by employing a statistical-based algorithm. In particular, the position of the experimental points are compared to the reference, simulated by the Radiative Transfer model.

In this work i-φ-MaLe takes advantage of R_{app} spectra, acquired by high-resolution spectrometers (Cogliati et al., 2019; Mohammed et al., 2019) in the 650-780 nm spectral region since it contains the amount of information necessary to estimate F, F_{RC} and the investigated biophysical parameters. Indeed, in a general perspective, in order to successfully apply i-φ-MaLe, the exploited spectral region has to include the wavelength intervals where variations in the analysed spectra shapes are caused by the retrieved parameters.

Architecture

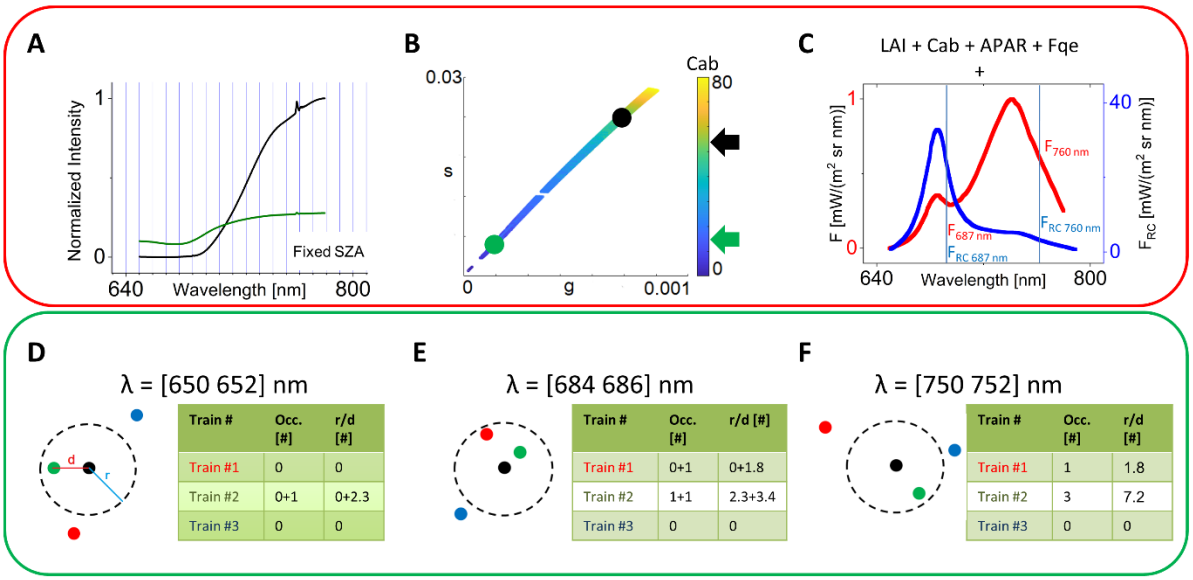

Figure 13. Schematic of the i-φ-MaLe method. *The Training and Retrieving procedures are framed by a red and a green line, respectively. (**A**) R_{app} spectra (dense/sparse vegetation in black/green, respectively) are separated in equally spaced spectral windows and transformed in points lying on the phasor plane by means of **Eq. 29**, as shown in panel (**B**). A supervised ML algorithm associates the parameters values to the coordinates of each point in the phasor plane, shown here color-coded for Cab. This procedure is repeated for each spectral window and therefore for each phasor plane. (**C**) Parameters retrieved by the algorithm. Panels (**D**)-(**F**) sketch the retrieving procedure for three 2-nm spectral windows. The algorithm assigns to the unknown test point (in black) the biophysical and F metrics values corresponding to the trained point with the highest occurrence and distance scores, as detailed in the text.*

The i-φ-MaLe pipeline is composed of two phases: a training (T) and a retrieving (R) procedures, which are shown in **Figure 13** framed by a red and a green line, respectively. Moreover, a flow chart of the method is provided in **Figure 14**: in this case, the pre-processing, training and test parts are color-coded in blue, orange and green, respectively. Training and test set are highlighted with red and dark green colors.

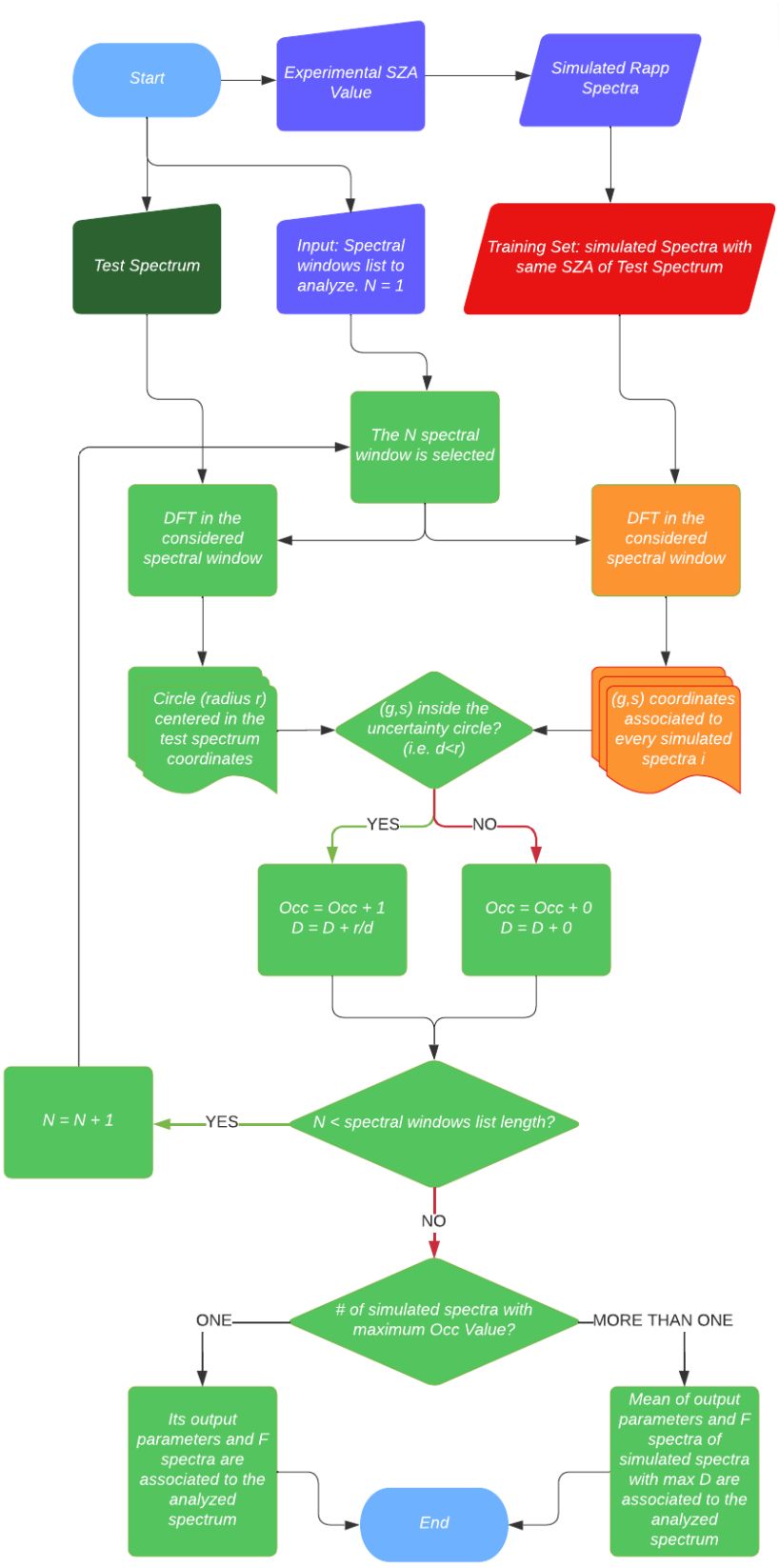

Figure 14. i-φ-MaLe method flowchart*. Blue-colored blocks represent the pre-processing pipeline steps, while the training and retrieval procedure are color-coded in orange and green, respectively. The training and test dataset are highlighted by red and dark green colors.*

During the training phase (T), a dataset of synthetic R_{app} spectra, created by coupling SCOPE (Soil Canopy Observation, Photochemistry and Energy fluxes - van der Tol et al.

2014, van der Tol et al. 2009) and MODTRAN5 (MODerate resolution atmospheric TRANsmission - Berk et al., 2011) Radiative Transfer codes, has been exploited. In particular, each R_{app} spectrum of the training set has been simulated by varying the parameters reported in **Table 1**.

Variable Name	Variable description	Range
SZA	Solar Zenith Angle [°]	20°-60° Step 2°
LAI	Leaf Area Index [m^2m^{-2}]	1-8
Cab	Chlorophyll a and b [$\mu g/cm^2$]	1-80
Fqe	Fluorescence Quantum Efficiency	0.010-0.024 Step 0.001

Table 1. Variability range of the MODTRAN+SCOPE input parameters*. The intervals cover the experimental variability characterizing both the daily and the annual trends of the investigated variables.*

The training phase can be summarized as follows:

1. The entire training dataset is sub-sampled by considering only the R_{app} spectra associated to the same solar zenith angle (SZA) of the test (or experimental) R_{app} analyzed by the algorithm during the retrieving phase;
2. Each spectrum is divided into equally spaced spectral windows (**Figure 13A**), whose width has to be larger than the sensor spectral resolution but small enough to prevent losing information on the spectral shape. For a fixed spectral resolution, this condition can be achieved by modifying the spectral windows amplitude and superposition. The accuracy of i-φ-MaLe has been quantified in dependence of these hyperparameters by simulating the spectral sampling interval (SSI) of two different spectrometers: QEPro (**Figure 18A** and **Figure 18B** – SSI = 0.15nm) and FLAME-S (**Figure 19A** and **Figure 19B** – SSI = 0.60nm). This analysis highlighted that, for an optimal algorithm configuration, each window is

expected to contain ~10 points of the R_{app} spectrum. During the analysis of experimental data (**Section 5.3**), 2 nm width spectral windows have been chosen since they guarantee the best retrieving performance, as reported in **Figure 18**.

3. In each window, the Discrete Fourier Transform (DFT) projects the R_{app} spectrum into a point lying in the phasor plane. The phasor coordinates (g, s) of the point are represented by the real and the imaginary parts of the first harmonic DFT of the normalized R_{app} spectrum, computed according to the following:

$$g\,(k) = \frac{\sum_{\lambda_s}^{\lambda_f} R_{app}(\lambda) cos(\omega k \lambda)}{\sum_{\lambda_s}^{\lambda_f} R_{app}(\lambda)} \qquad s\,(k) = \frac{\sum_{\lambda_s}^{\lambda_f} R_{app}(\lambda) sin(\omega k \lambda)}{\sum_{\lambda_s}^{\lambda_f} R_{app}(\lambda)} \qquad (26)$$

where λ_s and λ_f are the starting and final wavelengths related to each considered spectral interval, $\omega = 2\pi/\tau s$ with τs = number of spectral channels in each spectral interval, and k is the harmonic (k = 1 in our case). By computing the DFT of every R_{app} of the training dataset in the same spectral window, each spectrum is transformed in a point lying on the phasor space which is associated to Cab, LAI, APAR, F, F_{RC} and Fqe values characterizing the R_{app} spectrum. **Figure 13B** reports the points distribution in a selected phasor space, where each point is color-coded for the associated Cab value: this highlights that, generally, close points on the phasor plot are associated to similar parameter values. Unfortunately, points characterized by really different parameter values could lie superimposed due to the effect of a possible ill-pose problem among the investigated variables. This phenomenon is mitigated by iterating the projection procedure for every spectral window, obtaining multiple distributions of points characterized by different mutual relative positions in each phasor space. In this way, points superimposed due to the ill-pose problem will lie far apart in other phasor plots, while the position of points characterized by similar parameter values will be similar during every iteration.

The retrieval phase has been conceived to estimate F, F_{RC} and the biophysical parameters starting from an experimental R_{app} spectrum, which is not part of the training dataset. In particular, the user is able to set three different configurations of the retrieval phase:

ALG1 (standard): the entire test spectrum is divided into the same equally spaced spectral intervals exploited to separate the training set spectra. In the limit case of ideal data (no noise), the test point will lie very close or superimposed to the training point associated to the most similar parameter values in every phasor space. Since noise affecting the test spectrum scatters the points on the phasor plot around the ideal position (**Section 1.2**), i-φ-MaLe has been designed to automatically take into account the signal to noise ratio (SNR) by applying the following retrieval procedure:

1. A circle with radius r, centered around the test point (black dot in **Figures 13D-F**), is considered in each phasor space.

   When real acquired data are analyzed, the size of the uncertainty circle is automatically set by the algorithm by considering the noise characterizing the analyzed spectra. In particular, the radius r scales inversely with the signal to noise ratio value (Cutrale et al. 2017).
2. Each training dataset point lying into the circle is associated to two different scores, the Occurrence score Occ and the Distance score D, computed as:

$$Occ. = \sum_{i=1}^{N} \alpha_i \quad \& \quad D = \sum_{i=1}^{N} \left(\frac{r}{d_i}\right) \alpha_i \qquad (27)$$

$$where \quad \alpha_i = \{1 \quad if \quad d_i \leq r \; 0 \quad if \quad d_i > r$$

   Occ is the number of phasor planes where the training point lies within the circle of radius r, d_i is the Euclidean distance between the training and the test points (**Figures 13D-F**) and N is the number of the analyzed spectral windows.
3. The algorithm associates to the test spectrum the single parameter tuple P_T={LAI, Cab, APAR, Fqe, F and F_{RC} full spectra} corresponding to the training point with the highest Occ score. In case of tie, the algorithm selects the values corresponding to the nearest point

(highest D score, **Figure 13D-F**) among those with the same occurrence. If a tie is reached also for D, the algorithm computes their mean value. R_{app} spectra characterized by values of Cab < 5 μg/cm^2, LAI < 0.5 and APAR < 50 W/m^2 are automatically classified as soil.

ALG2 (best spectral windows): it relies on the application of the ALG1 procedure to retrieve independently each parameter by exploiting only the spectral windows characterized by the highest information content. As a consequence, if the same test spectrum is considered, the training point with the highest Occ score can be different for each retrieved variable. The best spectral windows related to each parameter have been selected by computing the Score Index S (**Table 2**) and they are compatible with the state-of-the-art theoretical expectation and experimental retrieving procedures. For instance, information concerning the parameters related to fluorescence emission (Fqe, F and F_{RC}) is almost completely provided by the O_2 bands, where the R_{app} spectrum is not influenced by Cab (indeed, it is usually retrieved in the 700-720nm interval).

Parameter	Spectral Interval [nm]
Cab	650-676, 682-686, 688-692, 698-758, 762-764, 766-780
LAI	650-662, 672-680, 694-716, 736-740, 748-758, 762-780
APAR	650-660, 672-680, 686-688, 692-780
Fqe	656-658, 674-680, 684-708, 714-770
F	672-680, 688-690, 694-780
F_{RC}	672-680, 688-690, 694-780

Table 2. Spectral intervals exploited by ALG2 *for the retrieving of biophysical and F parameters (Score Index > 0.25)*

To compute the Score Index S, for each test spectrum, the ALG1 procedure has been independently applied in each spectral window. If the retrieved parameter in the spectral

window corresponds to the true simulated value, the Score Index related to the single spectral window has been incremented S = +1, otherwise S = +0. This step has been repeated over 10000 R_{app} test spectra and normalized to 1 (where 1 denotes windows where the selected parameter is retrieved with the highest accuracy). In this way, a plot of S vs spectral windows (**Figure 15**) allows to highlight the spectral regions containing the highest information content to retrieve each parameter. A threshold S ≥ 0.25 was set to select the best spectral regions since this value assures many enough (>50) spectral windows to apply the statistical-based retrieval algorithm.

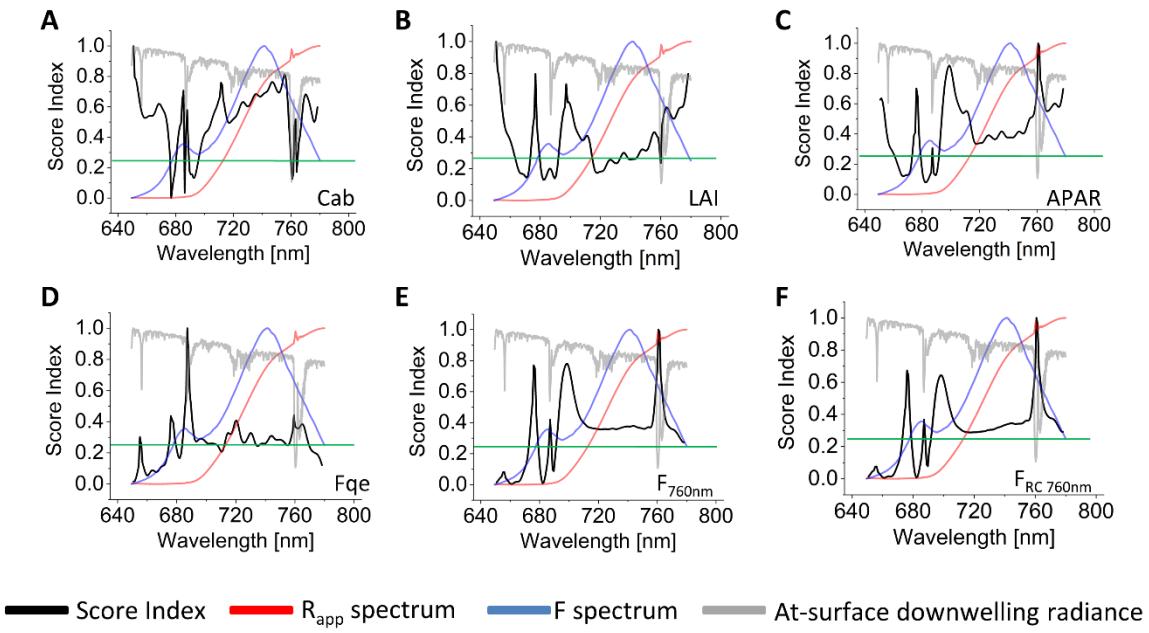

Figure 15. Score Index S, R_{app}, F and incoming radiance (normalized for visualization purposes) computed on sequential 2 nm spectral windows *for the following parameters: (A) Cab; (B) LAI; (C) APAR; (D) Fqe; (E) F at 760 nm; (F) F_{RC} at 760 nm. This trend is representative also for F and F_{RC} at 687 nm. The green line represents the Score Index threshold for the best windows selection.*

ALG3 (exclusion of oxygen bands): ALG1 has been optimized for the analysis of spectra acquired at tower-level (100-150 m from the canopy) by excluding the telluric oxygen absorption bands (i.e. O_2-A and O_2-B bands at 760 nm and 687 nm, respectively) since they are affected by reabsorption phenomena due to the telluric oxygen present in the atmosphere (**Section 2.2** for further information).

5.2 Method characterization and validation on simulated spectra

The training dataset exploited by i-φ-MaLe has been prepared by simulating 21 sets of 200.000 R_{app} spectra characterized by a different SZA value, which has been varied within the 20°-60° interval with a step of 2°. In each set, the LAI and Cab were varied between 1-8 m^2/m^2 and 1-80 $\mu g/cm^2$, respectively, while the variability due to the fluorescence contribution was obtained by changing Fqe between 0.010 and 0.024 with a step of 0.001. The canopy geometry was assumed spherical by fixing the Leaf Inclination Distribution Function (LIDF) parameters (LIDFa = -0.35 and LIDFb = -0.15). In particular, the spectra have been simulated by exploiting the same spectral sampling (SSI) of the instrument exploited to acquire the experimental data, which is a QEPro spectrometer (Ocean Insight, USA), characterized by a high spectral resolution in the wavelengths range in which the fluorescence occurs (650-800 nm), featured by a Full Width Half Maximum (FWHM) of 0.30 nm and a SSI = 0.15nm.

The results concerning the retrieval accuracy of ALG1 and ALG2 for a test set of 10000 spectra have been reported in **Table 3**.

Parameter	ALG1 - RRMSE (%) all spectral windows	ALG2 - RRMSE (%) selected spectral windows
Cab	1.0 ± 0.7	0.6 ± 0.3
LAI	2.0 ± 1.0	2.0 ± 1.0
APAR	0.7 ± 0.5	0.8 ± 0.6
Fqe	0.8 ± 0.1	0.7 ± 0.1
F (687 nm)	0.7 ± 0.7	0.7 ± 0.6
F (760 nm)	1.0 ± 0.8	1.0 ± 0.8
F_{RC} (687 nm)	0.9 ± 0.8	0.9 ± 0.7
F_{RC} (760)	0.9 ± 0.8	0.9 ± 0.7

Table 3. ***RRMSE (%) for the retrieval of biophysical parameters and F by exploiting algorithm ALG1 (whole spectral analysis) and algorithm ALG2 (parameter dependent spectral windows).*** *ALG2 shows slightly best results, in terms of accuracy, for the Cab parameter.*

The parameter whose retrieval is most affected by the choice of the algorithm is Cab, which is retrieved with half the initial error by ALG2. The fluorescence spectra are retrieved with a slightly higher error for the far red band with respect to the values in the red counterparts, while the fluorescence metrics at the photosystem level show the same RRMSE (%) values. Since slightly more accurate results are obtained by exploiting only selected windows for the retrieving of each parameter, with also a simultaneous half reduction of the computation time, in the rest of the analysis ALG2 has been therefore applied, except for data acquired from towers. A more refined analysis for ALG2 is presented in **Figure 16**, where the test set has been divided in steps of 5 µg/cm^2 for Cab, 0.5 m^2/ m^2 for LAI, 50 W/m^2 for APAR, and discrete intervals with a 0.001 width for Fqe. This data visualization approach is able to highlight, for each parameter, if the retrieval accuracy is uniform in the entire range in which the variable can be estimated by the algorithm.

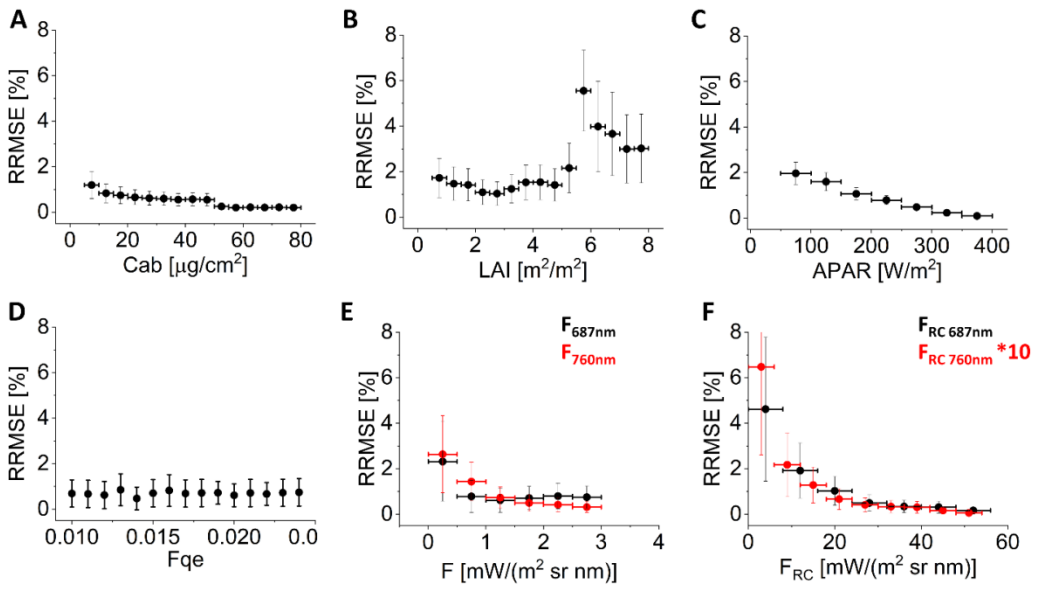

Figure 16. RRMSE (%) in dependence of equally spaced parameter intervals. *Panels refer to: (**A**) Cab, separated into 5 µg/cm^2 intervals; (**B**) LAI, divided in 0.5 m^2/m^2 steps; (**C**) APAR, separated into 50 W/m^2 intervals; (**D**) Fqe, divided into discrete 0.001 steps; (**E**) F at 687 nm and at 760 nm, separated into 0.5 mW/ m^2 sr nm intervals; (**F**) F_{RC} evaluated at 687 nm, divided in 8 mW/ m^2 nm steps, and F_{RC} at 760 nm, separated in 0.6 mW/ m^2 nm intervals. Parameters are retrieved by exploiting the algorithm ALG2.*

Retrieval accuracy is almost constant for Cab and Fqe (**Figure 16A** and **Figure 16D**, respectively), while APAR, F and F_{RC} are affected by higher uncertainties for low parameter values. The RRMSE referred to the LAI parameter is almost constant in the range [0 6], while the error associated to the parameter retrieval is higher for LAI > 6. This is due to the SCOPE radiative transfer model, which is less accurate in the spectra reconstruction for high LAI values.

Since the results concerning F and F_{RC} have been reported only at 687nm and 760nm, the capability of i-*φ*-MaLe to accurately retrieve the entire F and F_{RC} spectra has been highlighted in **Figure 17**, which shows a comparison between the reference and the retrieved F and F_{RC} spectra for 12 different scenarios.

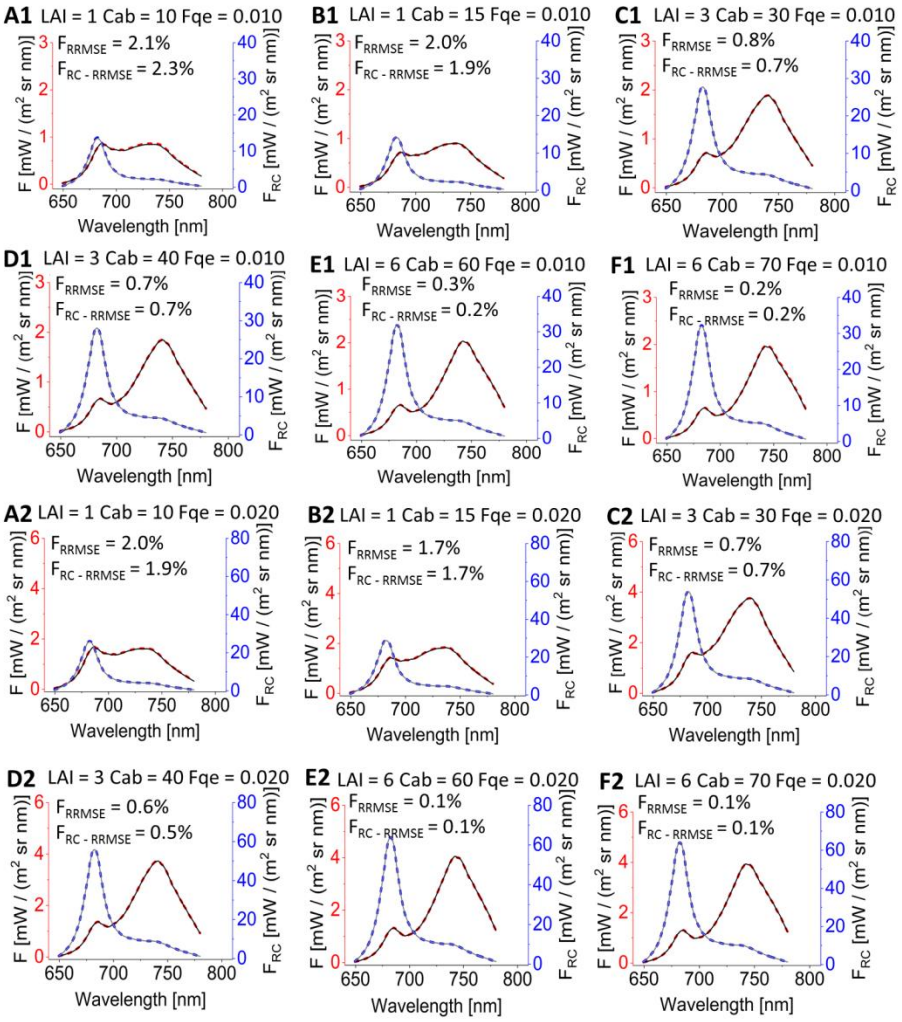

Figure 17. Comparison between F *(continuous black line)* ***and F_{RC}*** *(continuous gray line) retrieved spectra with the reference (thick dashed red and thick dashed blue lines,*

respectively) for selected simulated cases, as indicated in each graph with the corresponding RRMSE values.

In particular, it is demonstrated that lower intensity values of F and F_{RC} are affected by higher uncertainties.

In order to provide a more complete characterization of the method, the retrieval accuracy has been investigated by varying its hyperparameters: the windows width (**Figure 18A**), and the windows overlap (**Figure 18B**). Moreover, the possible presence of discrepancies due to different SZA have been investigated in term of algorithm accuracy (**Figure 18C**).

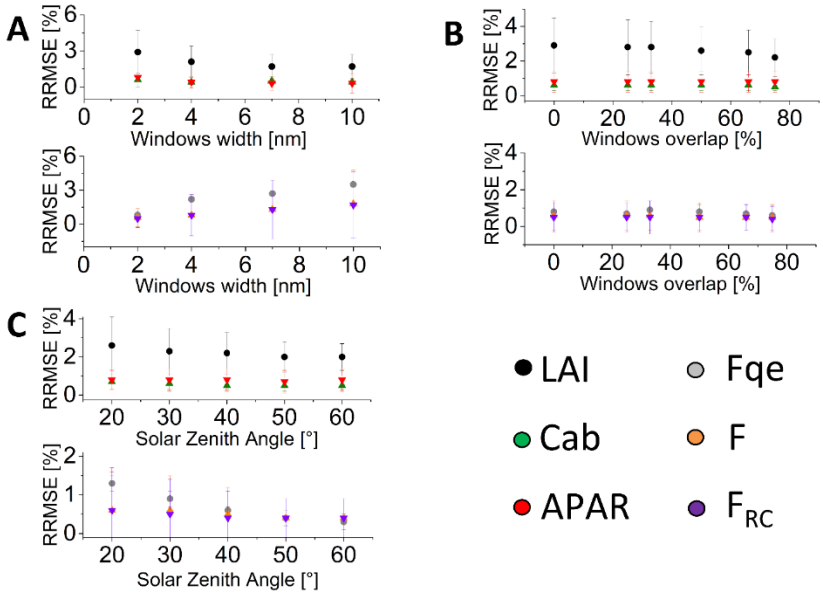

***Figure 18. RRMSE for the retrieving of Cab, LAI, APAR, Fqe, F and F_{RC} in dependence of**: (**A**) spectral windows width (SZA=40°); (**B**) spectral windows overlap (SZA=40°); (**C**) solar zenith angle values.*

The accuracy is almost unaltered by varying both the spectral windows overlap or the solar zenith angle values. Concerning the spectral windows width, a slightly better performance is obtained for Fqe, F and F_{RC} over narrower intervals.

In order to investigate the influence of the spectrometer properties on the parameters retrieval error in dependence of the hyperparameters, we repeated the same analysis by simulating the R_{app} spectra with a SSI=0.6 nm (typical of the FLAME-S spectrometer). The results have been reported in **Figure 19.**

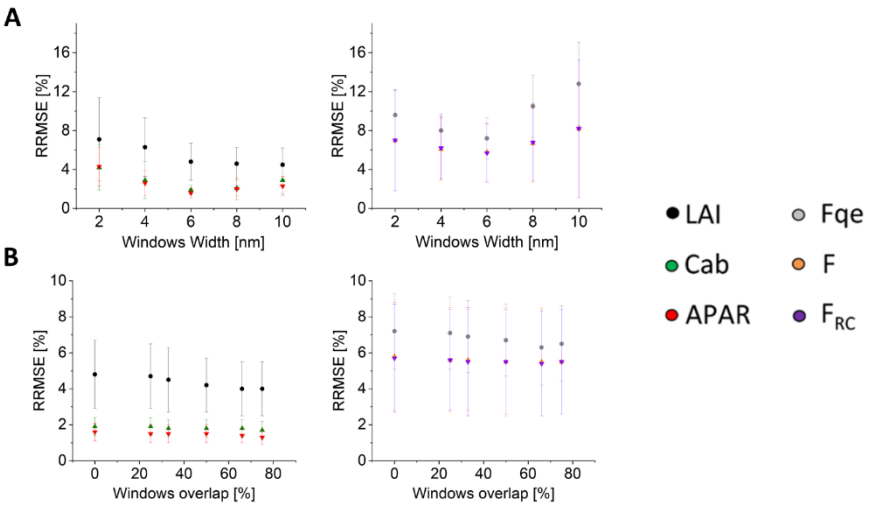

***Figure 19. RRMSE in the retrieving of Cab, LAI, APAR, Fqe, F and F_{RC} in dependence of the** (**A**) spectral windows width (SZA=40°) and (**B**) spectral windows overlap (SZA=40°) for Rapp simulated by exploiting the FLAME-S spectrometer properties.*

The results obtained in the FLAME-S scenario are compatible with the simulations performed to analyze the spectra acquired by QEPro since, also in this case, the accuracies are almost unaltered for variations of the spectral windows overlap. As expected, due to the lower spectral resolution, the overall accuracy of i-φ-MaLe decreases, in particular regarding the retrieval of Fqe, F and F_{RC}. The windows width characterized by the highest accuracy is 6nm (**Figure 19A**). Interestingly, the ratio between the best window width and the instrument SSI related to both QEPro (13) and FLAME-s (10) is similar, suggesting that, for an optimal algorithm configuration, each spectral window is supposed to contain ~10 experimental points to obtain the best results.

Finally, in order to maximize the algorithm efficiency in terms of computational costs, the same test dataset of 10000 spectra (where SZA = 40°) has been analyzed by exploiting a different number of simulations in the training dataset. The results are provided in **Figure 20**.

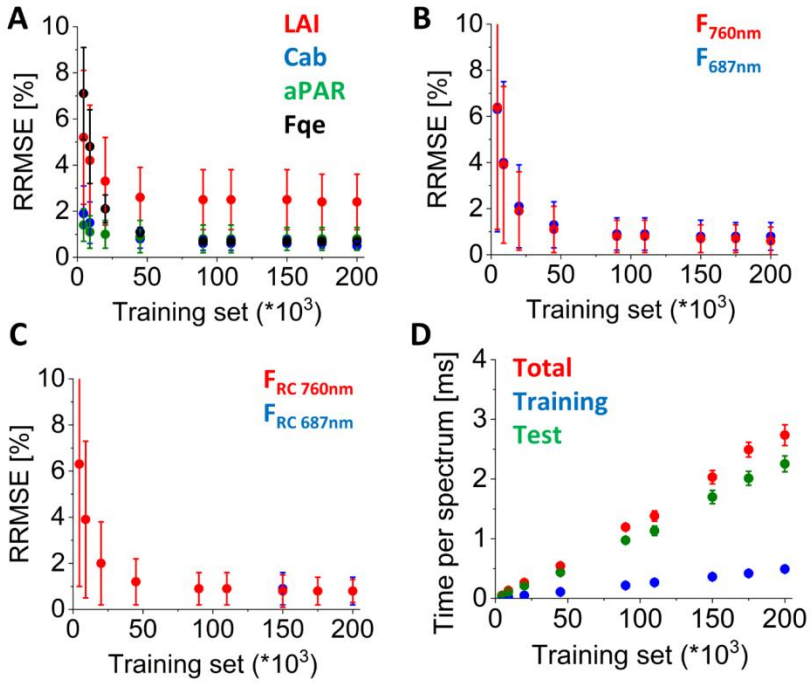

Figure 20. RRMSE computed in dependence of the elements in the training set with a varying Fqe. *The test dataset is characterized by 10000 points for each Fqe in the range 0.01-0.024 (steps of 0.001) and a total of 21 SZA values. Data refer to: (**A**) Cab, LAI, APAR, Fqe; (**B**) F at the O_2-A and O_2-B bands; (**C**) F_{RC} at the O_2-A and O_2-B bands. Panel (**D**) reports the processing time per spectrum (total= training + test of 10000 randomly extracted spectra with a 5-fold cross-validation).*

As expected, the RRMSE decreases while increasing the dataset size for all retrieved parameters, staying below the 7%. Moreover, for a training dataset size > 75000, the RRMSE associated to the retrieval accuracy is almost unaltered for every considered parameter. Furthermore, even though the computational time increases linearly with the number of training elements in the dataset (**Figure 20D**), for a training dataset size of 150000 spectra the time to process a single R_{app} spectrum in only 2ms. Since the QEPro spectrometer acquires a single spectrum in 1-2 minutes, the proposed algorithm assures a real time evaluation of the experimental data. Therefore, in a very limited computational time, i-φ-MaLe is able to retrieve F, F_{RC} and the biophysical parameters over a wide range, covering both the vegetation daily and seasonal variations (**Section 5.3** for further information).

5.3 Experimental data analysis

In this section, i-φ-MaLe has been applied in different experimental scenarios, from top of canopy to tower-level measurements.

5.3.1 Experimental data acquisition

At first, the algorithm has been exploited to analyze real data acquired at top of canopy level and assess the seasonal and diurnal trends of each variable characterizing two crops (forage and Alfalfa). These spectral measurements were acquired across the 2018 season on agricultural area in central Italy, i.e. forage (Feb 21st - May 24th) and alfalfa (May 25th - Jul 12nd) at 1-3 meters of height between the 07:00 up to the 17:00 (UTC time). In order to assess the seasonal and daily trends, 15 clear sky days have been selected by exploiting the illumination stability criteria (Cogliati et al. 2015a) and analyzed by i-φ-MaLe since the RT simulations have been designed to accurately describe this scenario. In particular, 4 selected full daily trends have been reported and each data point is the average value computed from spectra acquired within 30 minutes time intervals with about 1 minute sampling time and singularly analyzed by i-φ-MaLe. Concerning the seasonal trends, the 15 reported points have been obtained by averaging retrievals within 30 minutes around the solar noon.

Then, seasonal and daily sets of R_{app} spectra acquired at tower-level over a Downy oak forest in France have been analyzed in order to demonstrate the capability of i-φ-MaLe in providing consistent results also in presence of atmospheric oxygen reabsorption effects which modify the R_{app} shape around 687 nm and 760 nm. In this case, the spectral measurements have been collected from high towers (distance between the sensor and the canopy around 100 m) from March to October 2018. In particular, 8 clear sky days have been selected to evaluate the seasonal trends, while the full daily trends of two DOYs (145 and 196) have been provided. During both experimental scenarios, the R_{app} spectra have been acquired by means of the FLoX setup, equipped with two different spectrometers: the

QEPro (Ocean Insight, USA), which is characterized by a high spectral resolution in the wavelengths range of fluorescence (650-800 nm), a Full Width Half Maximum (FWHM) of 0.30 nm, a SNR of 1000 and a spectral sampling interval (SSI) of 0.15 nm, and the FLAME-s (Ocean Insight, USA), which covers the VIS-NIR interval, with a FWHM of 1.7 nm, a SNR of 250 and SSI of 0.6 nm. The proposed algorithm exploits only the QEPro spectrometer since our aim is demonstrating that the spectral information in the range 650-800nm is enough to retrieve biophysical parameters and fluorescence spectra. Further details about the FLoX setup are provided by Cogliati and colleagues (Cogliati et al. 2019).

5.3.2 Seasonal trends of experimental measurements at canopy level

In this section, the ALG2 configuration has been exploited to analyze experimental data acquired at canopy-level (1-3 m between the top of canopy and the spectrometer). The noise affecting the R_{app} spectra (SNR ~ 500) has been also considered by the algorithm. The final aim of this analysis is evaluating the consistency, over the growing season, of the biophysical and F parameters retrieved by i-φ-MaLe for two agricultural crops (forage and alfalfa).

Figure 21 shows the first performed analysis, which is the comparison between the F_{687} and F_{760} parameters retrieved by applying the ALG2 configuration and the corresponding values provided by other two already published algorithms: SFM (Meroni et al. 2010, Mazzoni et al. 2010, Cogliati et al. 2015b, **Figure 21A**) and SpecFit (Cogliati et al. 2019, **Figure 21B**). This analysis highlights the good agreement among the considered methods, in particular for early growth stages. The F_{687} values retrieved by the proposed algorithm, if compared to SFM (**Figure 21A**), are characterized by a slope estimated by standard least square linear regression model of 1.23, slightly above to the 1:1 trend, while F_{760} is characterized by a slope of 0.98, closer to 1. This discrepancy in the obtained F_{687} and F_{760} values is probably related to the different retrieving pipelines characterizing i-φ-MaLe and SFM: the F

retrievals in the oxygen bands by SFM are independent, while i-φ-MaLe extracts a single F spectrum from which F_{687} and F_{760} are obtained in a dependent manner.

The comparison between i-φ-MaLe and SpecFit is provided in **Figure 21C-F**. In particular, exemplary F spectra extracted from the same experimental R_{app} and their related variability range are shown both for i-φ-MaLe (in red) and SpecFit (in black). The overlap between the F spectra demonstrates that i-φ-MaLe and SpecFit provide compatible results in each analyzed spectral window. Slight discrepancies concerning their shape are detected only for F spectra characterized by really low values of LAI and Cab (LAI < 1 m^2/m^2, Cab < 10 $\mu g/cm^2$) (**Figure 21C**).

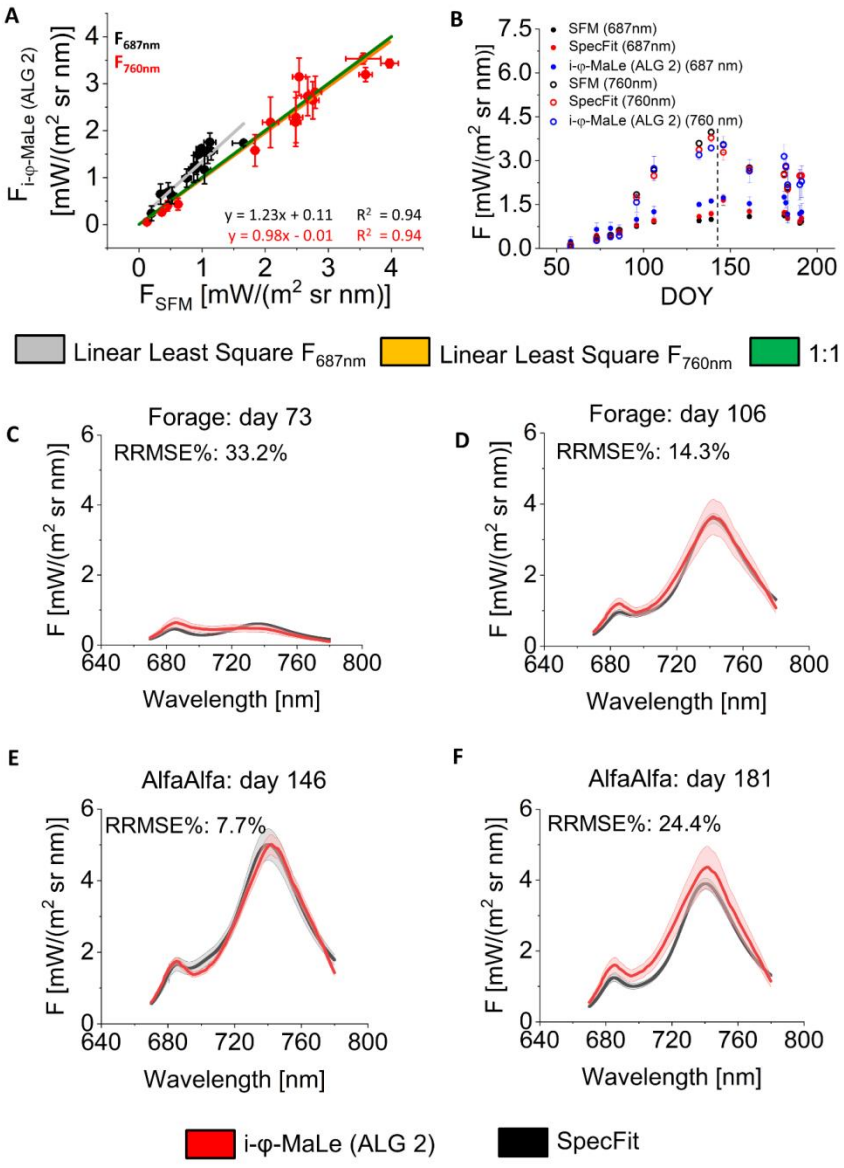

Figure 21. Comparison of the experimental data analysis results provided by SFM, SpecFit and i-φ-MaLe. *(**A**) Scatter-plot between F_{687} and F_{760} metrics obtained through algorithm ALG2 and retrieved by SFM. (**B**) Comparison of the F_{687} and F_{760} metrics retrieved by means of algorithm ALG2 and by SFM and SpecFit methods. Data are related*

*to both forage (Day Of the Year, DOY=58-139) and alfalfa (DOY=146-191) crops, separated by the black dashed line. **(C)-(F)** Comparison of the full fluorescence spectra retrieved by means of our method (red) and through SpecFit (black). The data reported in panels **A** and **B** represent the average and the related range of variability of 40 spectra, acquired for a total of 30 minutes around noon in clear sky conditions. The integrated RRMSE (%) is computed by a point-by-point comparison of the retrieved fluorescence spectra.*

In order to verify if the exclusion of the oxygen bands is able to considerably affect the retrieval, the seasonal trends of the biophysical parameters and fluorescence signals have been reported in **Figure 22** both for the ALG2 and ALG3 configurations.

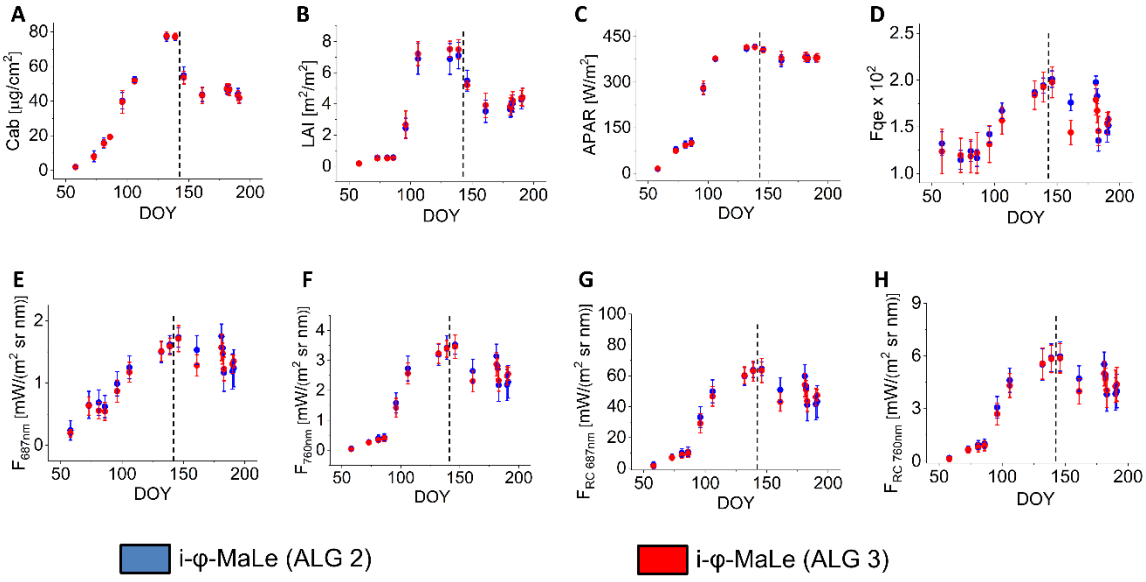

***Figure 22. Seasonal trends analysis of top of canopy-level measurements.** The seasonal variations of **(A)** Cab, **(B)** LAI, **(C)** APAR, **(D)** Fqe, **(E)** F at 687 nm (O_2-B band), **(F)** F at 760 nm (O_2-A band), **(G)** F_{RC} at 687 nm and **(H)** F_{RC} at 760 nm, retrieved through i-φ-MaLe by exploiting the best windows condition (blue, ALG2 configuration) and by excluding the O_2 bands (red, ALG3 configuration), are reported. In order to provide a simple and efficient representation of the seasonal trends, each data point is the average value among 40 spectra acquired within 30 minutes around solar noon in clear sky conditions for both forage (DOY=58-139) and alfalfa (DOY=146-191) crops, separated by the black dashed line. The errors associated to each data point consider both the mean standard deviation computed among the considered spectra and the retrieving error of the proposed algorithm.*

The biophysical parameters retrieved by the two algorithms are consistent with the known behaviors of these types of crops and coherent with the results obtained from the simulated dataset. The Fqe shows lower values until DOY=86, when it increases until the end of the growth season. Moreover, during the whole seasonal trend, F_{687} is characterized by lower

values compared to F_{760}. This behavior is compatible with the results reported by SpecFit (Cogliati et al. 2019). Fluorescence signals retrieved by both algorithm configurations show comparable values and, at the end of the growing season, the far-red F is significantly higher than the red fluorescence. Overall, this trend is consistent with the increasing canopy chlorophyll content along the growing season, causing a larger re-absorption in the red wavelength range and a gradual increase of far-red F. The fluorescence signal at the photosystem level, F_{RC}, observed in **Figure 22,** shows a behavior in agreement with the lack of re-absorption processes, implying that both F_{RC687} and F_{RC760} follow the same trend with $F_{RC687}>F_{RC760}$. This result is coherent with other published studies (Van der Tol et al. 2019).

For alfalfa, the drop at DOY = 161 is due to harvesting and we can observe that the Fqe and the fluorescence at the canopy and at the photosystem levels decrease after the cut and then start increasing during the second crop growth-cycle, although none of the parameters recover their initial values. Moreover, all the parameters retrieved by means of ALG2 or ALG3 are characterized by a compatible range of values.

This reflects that the information related to F is contained also in spectral windows other than the oxygen bands, as already highlighted in **Figure 15**. This result also demonstrates that the reabsorption effect of the atmosphere in the O_2 bands is negligible for spectra acquired at top of canopy level. A quantitative evaluation of this statement is reported in **Figure 23**, where a high correlation among the parameters estimated by ALG2 and ALG3 configurations is shown.

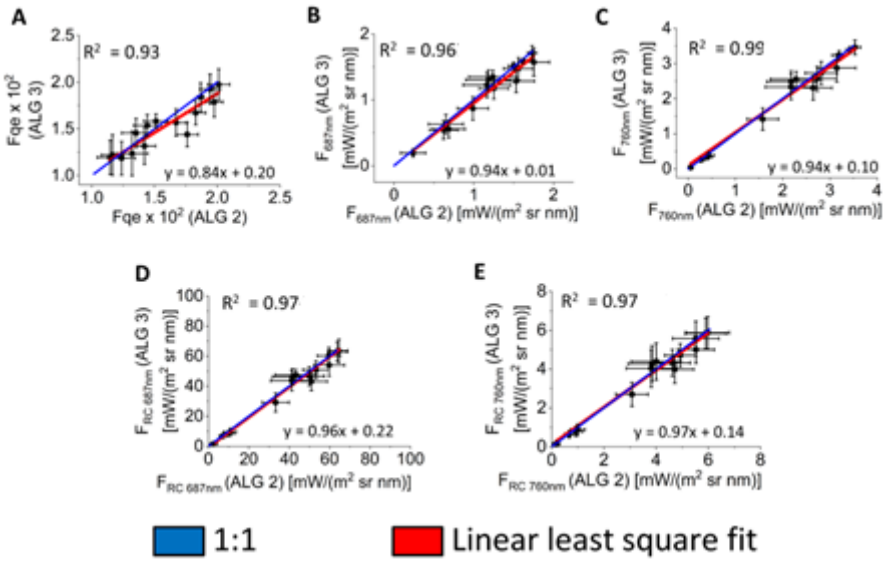

Figure 23. Scatter plots between values retrieved by our algorithm excluding the O_2 bands (ALG3) and by exploiting the best selected windows (ALG2)*, referring to: (A) Fqe; (B) F at 687 nm (O_2-B band); (C) F at 760 nm (O_2-A band); (D) F_{RC} at 687 nm (O_2-B band); (E) F_{RC} at 760 nm (O_2-A band). The red line is the linear least square fit, whereas the blue line refers to the 1:1.*

A further analysis can be provided by characterizing the correlations between F and the involved biophysical parameters (**Figure 24**).

The fluorescence signals at 687 nm and at 760 nm at both canopy and at the photosystem levels clearly vary in relation to the biophysical parameters. Their trends are compatible with the literature (Yang et al., 2015; Yang et al., 2017), further demonstrating the consistency of the values retrieved by *i-φ-MaLe* during the experimental data analysis. Fluorescence signals (F and F_{RC}) show different trends in dependence of the chlorophyll content (Cab, LAI and APAR). F dependence on Cab (**Figure 24, A-D**) and LAI (**Figure 24, E-H**) follows a more or less linear trend modulated by saturation effects for Cab >60 μg/cm^2 and LAI > 5 m^2/ m^2. F and F_{RC} retrieved from experimental data follow instead a linear trend as a function of APAR (**Figure 24, I-L**).

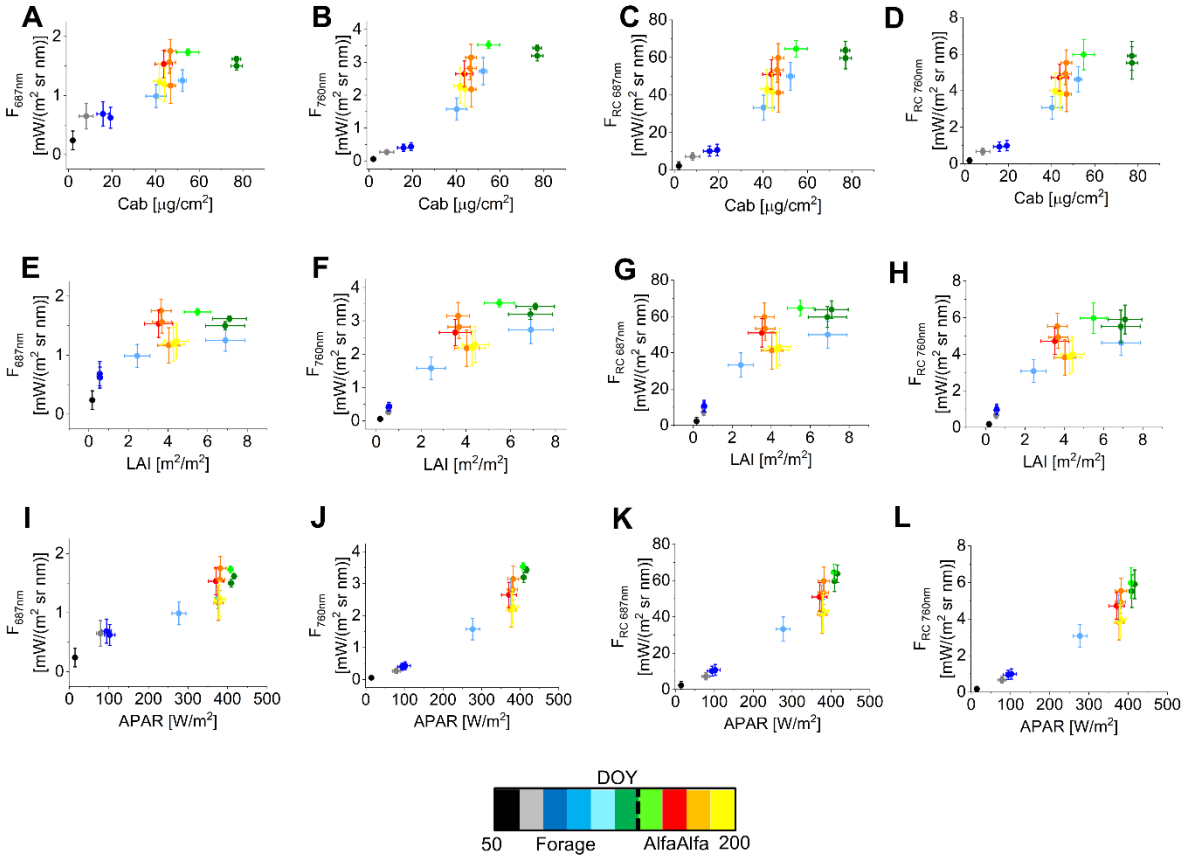

Figure 24. Comparison of the estimated experimental seasonal fluorescence metrics in dependence of Cab, LAI and APAR. *Panels (**A**)-(**D**) are related to the values of F at 687 nm, F at 760 nm, F_{RC} at 687 nm and F_{RC} at 760 nm respectively, in dependence of Cab. (**E**)-(**H**) refers to the same parameters reported as a function of the leaf area index, while panels (**I**)-(**L**) report the fluorescence metrics versus the APAR values. Points are color-coded for DOY and are related to forage (DOY<146) and alfalfa (DOY>146) crops. The black dashed line in the legend separates colors referring to forage and alfalfa crops.*

5.3.3 Daily trends of experimental measurements at canopy level

The daily evolution of F and the biophysical parameters for DOY = 73, 106, 139 (forage) and DOY=182 (alfalfa) is reported in **Figure 25**.

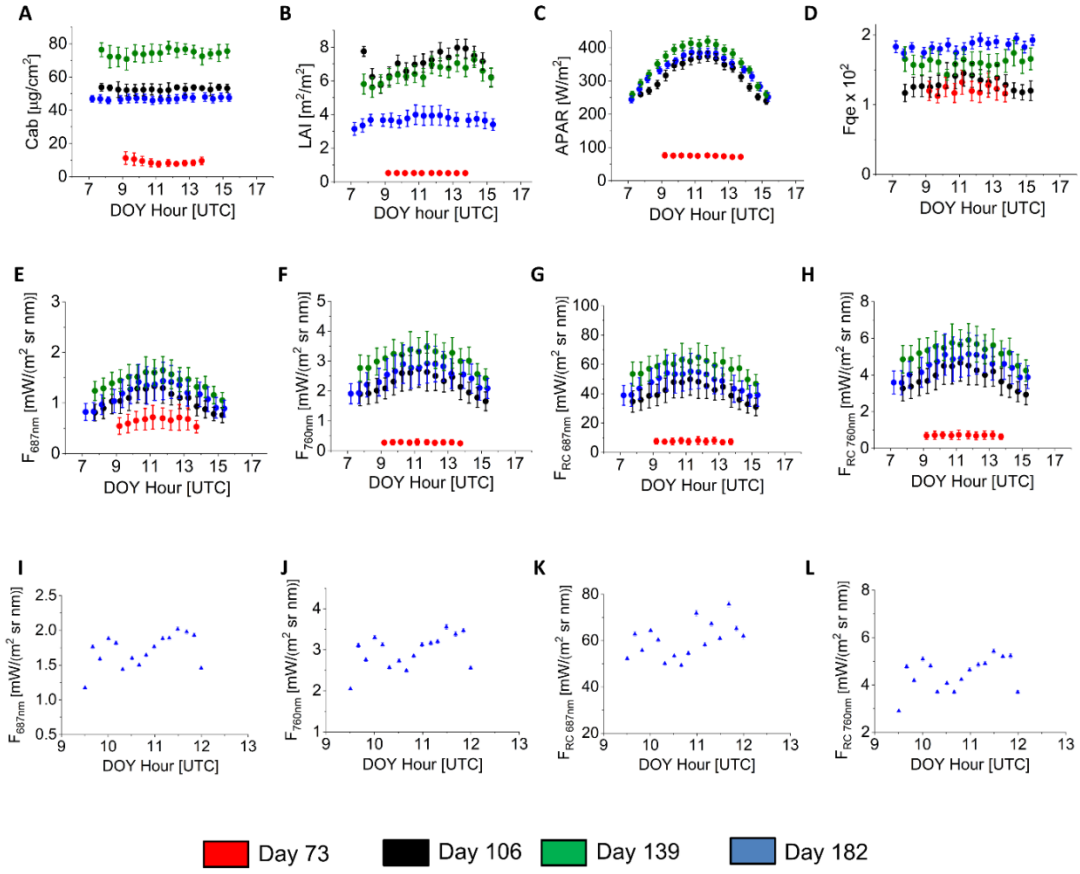

Figure 25. Daily trends analysis of top of canopy-level measurements. *Daily variations of (**A**) Cab, (**B**) LAI, (**C**) APAR, (**D**) Fqe, (**E**) F at 687 nm (O2-B band), (**F**) F at 760 nm (O2-A band), (**G**) F_{RC} at 687 nm and (**H**) F_{RC} at 760 nm, retrieved through ALG2 for different DOYs. Each data point is the average value among 40 spectra acquired within 30 minutes around each hour in clear sky conditions for both forage (DOY=73, 106, 139) and alfalfa (DOY=182) crops. Errors associated to each data point consider both the mean standard deviation computed among the considered spectra and the retrieving error of the proposed algorithm.* ***Panels I-L*** *report the daily trend of the fluorescence metrics from 9.30 to 12:00 UTC. Each data point refers to the analysis of a single reflectance spectrum with a 10 minutes sampling time and the associated error is due to the retrieval method for the experimental SNR=750.*

The retrieved Cab (**Figure 25A**) and LAI (**Figure 25B**) parameters are almost constant during the daily measurements. Less accurate results are associated to the LAI parameter for DOY 106 and 139 since SCOPE hardly manages LAI values in the [6 8] m^2/m^2 interval. APAR (**Figure 25C**) shows, instead, a diurnal variability with maximum values at low SZA (around solar noon). At morning and afternoon, the sun is lower than at midday and the penetration of sunrays is lower. Thus, the largest contribution of the sunlit foliage and of the lower leaf layers occurs at midday when the sunrays penetration is the highest. DOY 73

regards forage at the beginning of the growth (with the presence of a bare soil component, as observed from field spectral measurements) so that this trend is different from the others.

While the seasonal behavior of the F parameters primarily responds to the varying amount of chlorophyll in time, the daily response is largely modulated by photosynthetic activity, which is highest at noon and corresponds to the tiny positive parabolic trend (**Figure 25E-H**). We can also observe a subtle midday depression in the F signals (**Figure 25I-L**), particularly evident at DOY 182 in the alfalfa crops around 10.30 UTC (12.30 Local Mean Time - LMT), arising from photosynthetic activity (Campbell et al. 2019). The detection of the midday depression is a really interesting achievement since this phenomenon, which is a slight reduction of the vegetation fluorescence emission around the solar noon, highlights vegetation stress caused by an excess of energy coming from the sunlight. Also in this case we observe larger values of F_{RC} with respect to F, indicating the robustness of the algorithm. Moreover, the trends characterizing fluorescence values are similar to the results obtained by other already published methods (Campbell et al. 2019, Lu et al. 2020). A further indication of the reliability of the proposed method can be also provided by the stability of the fluorescence yield Fqe (**Figure 25D**), which, in absence of stress, should preserve similar values through the day. F is closely related to the light harvesting process and responds timely to rapid changes in photosynthesis (Baker, 2008). Photosynthesis is a process actively regulated by different mechanism according to the changes in the illumination occurring during the day (Romero et al. 2020, Pinto et al. 2020). Since photosynthesis and F compete for the same excitation energy, their efficiency is strictly linked to each other (Baker 2008). This could explain the small Fqe oscillation observed at diurnal scale, under not stressed conditions.

Finally we considered the Gross Primary Production (GPP), which is the total amount of carbon compounds produced by the photosynthesis of plants in an ecosystem in a given period of time. In particular, GPP is the golden standard parameter to monitor the photosynthetic process occurring in vegetation and depends on some of the biophysical parameters retrieved by the proposed algorithm (Rossini et al. 2012). Here, its daily trend

has been reported for DOY = 182, together with the APAR and F metrics (**Figure 26,** each considered variable has been normalized in the range [0 1] to better highlight the reciprocal behaviors).

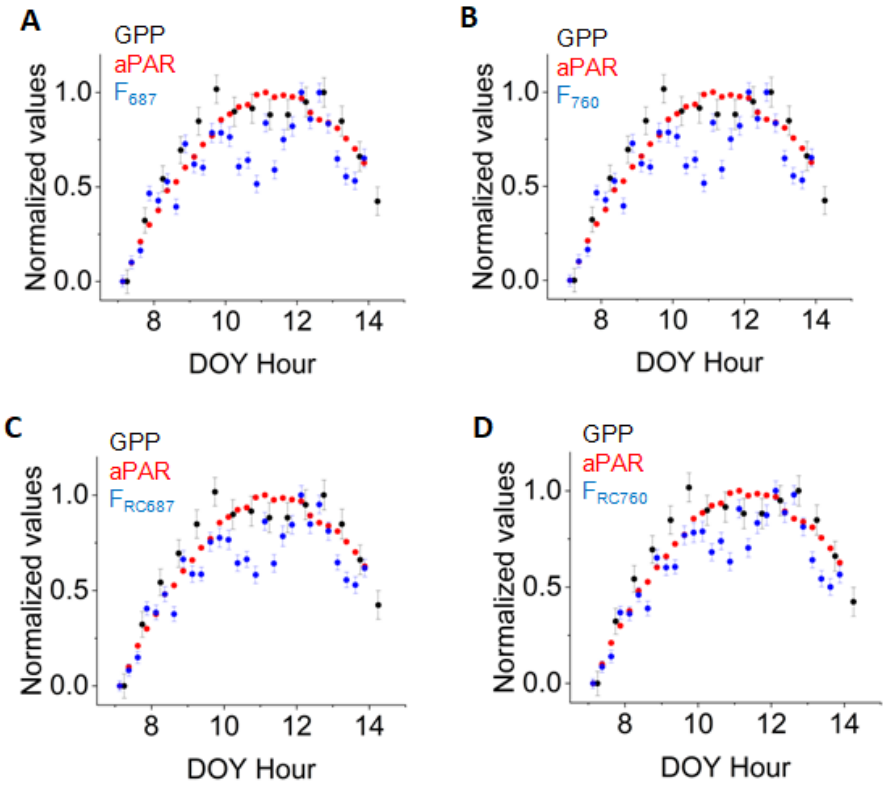

Figure 26. Daily variation of GPP *and APAR in comparison with* ***(A)*** F_{687}, ***(B)*** F_{760}, ***(C)*** F_{RC687}, ***(D)*** F_{RC760} *retrieved through the ALG2 algorithm for DOY = 182. Each data point is the average value among 40 spectra acquired within 30 minutes around each hour in clear sky conditions. Errors associated to each data point consider both the mean standard deviation computed among the considered spectra and the retrieving error of the proposed algorithm.*

As expected by the literature, the GPP trend is dependent on the retrieved APAR, (Rossini et al. 2012). This trend is unaffected by the presence of a midday depression of F around the solar noon. This further demonstrates that the different parameters are retrieved independently by the proposed algorithm and that their trends are compatible with the literature. Moreover, this analysis creates new insights about the possibility to exploit the proposed method to estimate new crucial variables able to provide a more complete description and quantification of the photosynthetic processes occurring in the vegetation, as the GPP.

5.3.4 Seasonal and Diurnal trends of Tower-level experimental measurements

In order to evaluate the consistency of the results obtained by exploiting the ALG3 configuration, i-φ-MaLe has been applied to retrieve the seasonal and daily trends characterizing a Downy Oak (*Quercus pubescens*) forest in France (43.931 N; 5.713 E). In tower-level acquisition conditions, atmospheric compensation techniques are needed to retrieve F at the O_2 bands. As a consequence, since SFM and SpecFit fluorescence retrievals are based on O_2 bands, a comparison between their outputs and ALG3 results is not available. From a preliminary analysis, it is interesting to highlight that the proposed method provides meaningful (i.e. positive) values without needing a proper atmospheric compensation differently from already developed algorithms based on O_2 bands information content. Moreover, by applying i-φ-MaLe, the complex and time-consuming application of atmospheric correction models devoted to modify the R_{app} spectra shape are not needed and no pre-processing procedures are necessary.

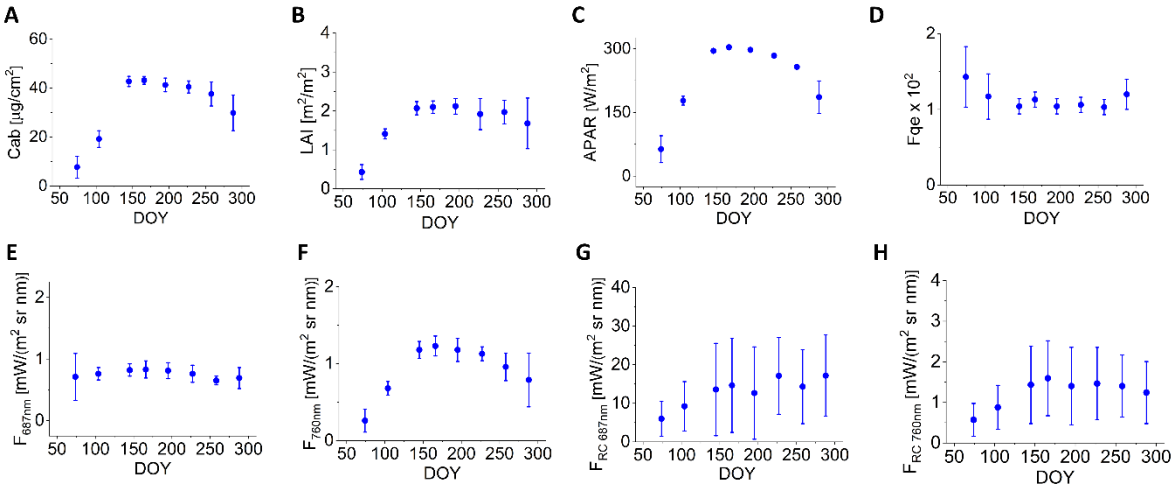

Figure 27. Seasonal trends analysis of tower-level measurements. *Seasonal variations of (**A**) Cab, (**B**) LAI, (**C**) APAR, (**D**) Fqe, (**E**) F at 687 nm (O_2-B band), (**F**) F at 760 nm (O_2-A band), (**G**) F_{RC} at 687 nm and (**H**) F_{RC} at 760 nm, retrieved through our algorithm by excluding the O_2 bands (ALG3 pipeline). In order to provide a simple and efficient representation of the seasonal trends, each data point is the average value among 40 spectra acquired within 30 minutes around solar noon in clear sky conditions for a deciduous forest of Downy oak in France. The errors associated to each data point consider both the mean standard deviation computed among the considered spectra and the retrieving error of the proposed algorithm.*

F and F_{RC} follow the APAR trend (**Figure 27C, E-H**), in agreement with the top of canopy-level measurements. Moreover, the F spectrum is characterized by higher values in the red with respect to the far-red until DOY=145, when an inversion of this trend is recorded. Instead, F_{RC} maintains a constant trend after reaching its maximum value at DOY=145, with $F_{RC687} > F_{RC760}$. It is noteworthy that the algorithm allows to retrieve the fluorescence emitted at the reaction center level with a reasonable accuracy at least for the annual variations. Even though we excluded the oxygen bands from the spectral range of the analysis, we still could give an estimate of the F_{RC} signals, although affected by large uncertainties, mainly related to the variability among consecutive acquired R_{app} spectra. Finally, Fqe starts from a value of 0.014±0.004 and then slightly decreases to 0.010±0.001 at DOY=145, maintaining a constant trend until DOY=258.

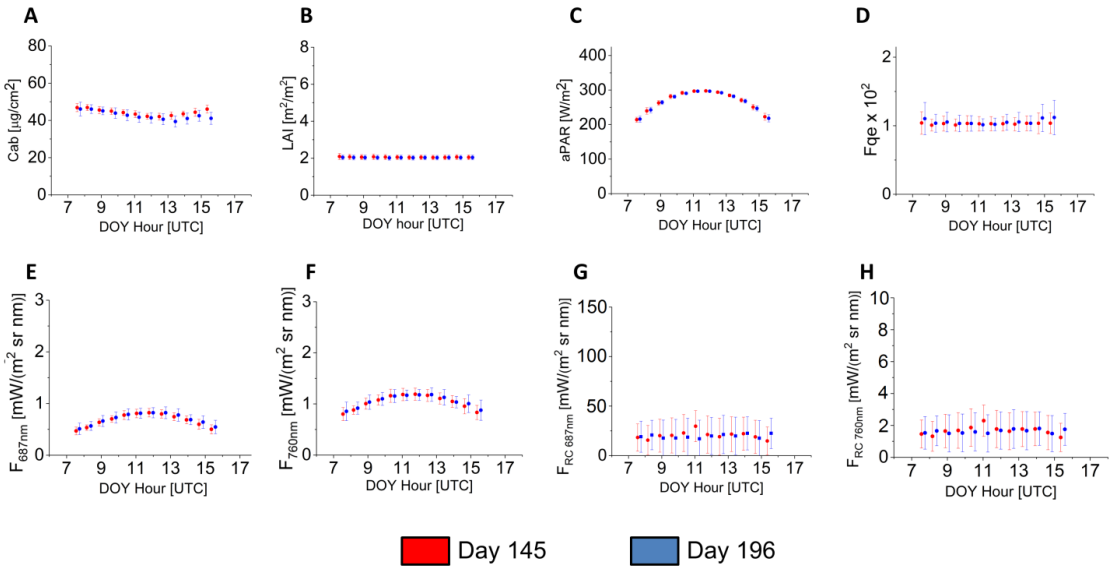

Figure 28. Daily trends analysis of tower-level measurements. *Daily variations of (**A**) Cab, (**B**) LAI, (**C**) APAR, (**D**) Fqe, (**E**) F at 687 nm (O2-B band), (**F**) F at 760 nm (O2-A band), (**G**) F_{RC} at 687 nm and (**H**) F_{RC} at 760 nm, retrieved through ALG3 configuration from tower-level measurements for different DOYs (Day 145 and Day 196). Each data point is the average value among 40 spectra, acquired within 30 minutes around each hour, in clear sky conditions. Errors associated to each data point consider both the mean standard deviation computed among the considered spectra and the retrieving error of the proposed algorithm.*

Also the daily trends of the biophysical and F parameters (**Figure 28**), retrieved from the tower data with the exclusion of the oxygen bands, are aligned with the outcomes of already published methods (Campbell et al., 2019, Lu et al., 2020) and with LAI and Cab values measured on the ground level, 1.90±0.71 m^2/m^2 and 45.6±5.8 $\mu g/cm^2$, respectively.

6. Φ-Norm, a phasor approach to normalize biopsies staining

Φ-Norm is a novel geometrical method conceived to provide a fast and accurate color normalization of stained tissue sections, supporting the application of Artificial Intelligence techniques to diagnosis automatization. The method is fully automated: consequently, its application does not require specific programming knowledge. Moreover, due its fully geometrical nature, Φ-Norm needs only a common PC, without any specific technical requirement available in every pathology laboratory (RAM = 16GB, CPU AMD A4-3420 APU). The algorithm, described in the following section, has been validated by normalizing the color content of H&E-stained human and murine liver tissue sections, available at the University of Leeds Virtual pathology Repository (ULVPR, https://archive.researchdata.leeds.ac.uk/) or acquired at the German Cancer Research Centre (DKFZ). The output normalized images have been quantitatively compared with the results obtained by other normalization methods (Reinhard et al. 2001, Macenko et al. 2009) in terms of 3D color distribution in the RGB colour-space. Then, the same analysis has been performed for tissue sections treated with a different staining, the PicroSirius Red (SR), in order to quantify and demonstrate the robustness of the proposed approach in performing accurate color normalization for different staining protocols. Moreover, a novel parameter, the Color Mean Distance (CMD), has been exploited to identify the most suitable method to reduce the color discrepancy among different tissue sections. Indeed, this is crucial to effectively apply machine learning techniques, which are highly affected by color discrepancies of the same biological structures in different samples.

6.1 Φ-Norm algorithm

Φ-Norm is a novel method conceived to perform stain normalization by exploiting the DFT geometrical properties in the phasor space. As extensively reported in literature (Fereidouini et al., 2012), if two different F_1 and F_2 spectra with the related coordinates on the phasor

plane (G_{F1},S_{F1}) and (G_{F2},S_{F2}) are considered, the coordinates (G_T,S_T) of the spectrum F_T = aF_1+bF_2, where a+b = 1 are defined as:

$$G_T = aG_{F1} + bG_{F2} \qquad (28)$$

$$S_T = aS_{F1} + bS_{F2} \qquad (29)$$

In Φ-Norm, this property has been applied to whole-slide images stored in RGB format (**Section 3.1**). In this context, the RGB spectrum of each pixel is represented by the X = [R,G,B] vector, where R, G and B represent the intensity value of the Red, Green and Blue channels.

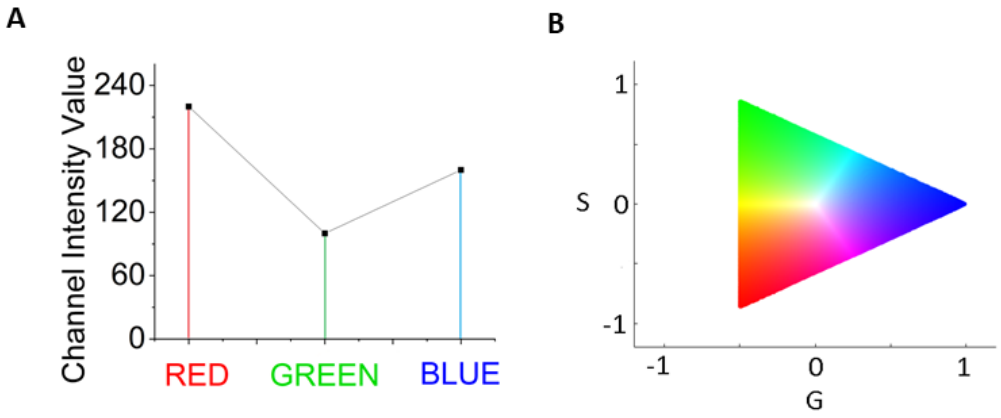

***Figure 29**: **Analysis of the RGB spectra projections on the phasor plane**. **A**: example of a RGB spectrum. In this case, the RGB values are [220,100,170]. **B**: scatter plot of the 2563 possible RGB spectra characterizing a 8-bit RGB format image, color-coded by their representative color.*

An example of RGB spectrum of a pixel characterized by the coordinates R=220, G=100, B=170 is shown in **Figure 29A**. Since a RGB spectrum can be considered as composed by the three spectra [R,0,0], [0,G,0] and [0,0,B], **Eq. 28** and **Eq. 29** are adapted as follows:

$$G_{[R,G,B]} = aG_{[R,0,0]} + bG_{[0,G,0]} + cG_{[0,0,B]} \qquad (30)$$

$$S_{[R,G,B]} = aS_{[R,0,0]} + bS_{[0,G,0]} + cS_{[0,0,B]} \qquad (31)$$

$$with \quad a + b + c = 1 \qquad (32)$$

In a standard 8-bit RGB format image (2^8 = 256 different intensity per channel), 256^3 (more than 16 million) different colors are represented, with their projections (**Figure 29B)** obtained by computing the DFT of each RGB spectrum X(n) = [R,G,B], as:

$$G_F = \frac{\sum_{n=1}^{3} X(n)cos(\frac{2\pi n}{3})}{\sum_{n=1}^{3} X(n)} \qquad S_F = \frac{\sum_{n=1}^{3} X(n)sin(\frac{2\pi n}{3})}{\sum_{n=1}^{3} X(n)} \qquad (33)$$

where n is the n-th color channel. In this phasor plot configuration, it can be demonstrated that if two spectra are projected in the same point, they are characterized by the same Hue and Saturation parameters (HSV colour-space), while their Value is undetermined. Indeed, the degeneration of the Value is caused by the normalization in **Eq. 33**: if two colors differ only for their Value parameter, the ratio between the intensity of each channel and the sum of the channels is the same, causing the degeneration.

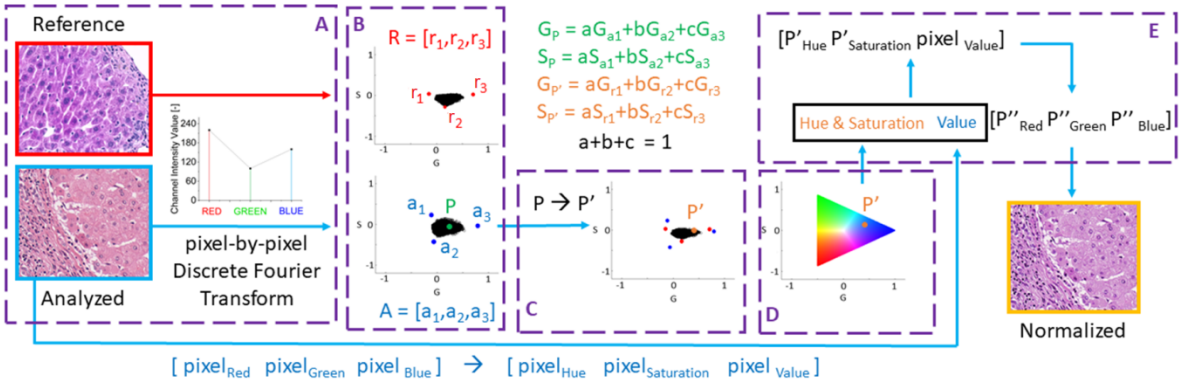

Figure 30: Schematic of the Φ-Norm method. ***A)*** *The RGB spectrum of each pixel is projected on the phasor plot for both the reference and the analyzed images.* ***B)*** *The respective staining vectors R and A are automatically retrieved by Φ-Norm, as they are the vertices of the best triangle inscribing all the pixel projections in the phasor plane. Each pixel projection of the analyzed image is mapped from the coordinates system expressed by the staining vectors to the system of the reference image by exploiting* ***Eq. 30, Eq. 31, Eq. 32****. The Hue and Saturation parameters of the color associated to the pixel projection in the new system are combined with the Value parameter obtained from the original image. Finally, the normalized color vector, lying in the HSV colour-space, is transformed in the RGB format.*

A scheme of the Φ-Norm method, which has been implemented in MATLAB, is reported in **Figure 30**. The proposed approach is unsupervised, fully automated and requires two input images: the analyzed image, whose staining needs to be normalized, and a reference image:

1) The algorithm acts both on the analyzed and reference images, projecting the RGB spectrum of each pixel on the related phasor plots by exploiting **Eq. 33** (**Figure 30A**).

2) For each phasor plot, the algorithm automatically identifies the best triangle inscribing all the projected points by exploiting a minimal bounding detection protocol, automatically implemented by MATLAB libraries. The triangle vertices are identified by the ordered triplets R = $[r_1,r_2,r_3]$ and A = $[a_1,a_2,a_3]$ for the reference and analyzed image, respectively (**Figure 30B**). These triplets represent the coordinates of the RGB spectra projection related to the pure contribute of a single staining in the image. For instance, in **Figure 30B**, a_3 and r_3 correspond to the pure contribute of Hematoxylin in the analyzed and reference image, respectively, while a_2 and r_2 represent the pure contribute of Eosin and a_1 and r_1 refer to the background. As a consequence, R and A correspond, in the phasor plane, to the staining vectors exploited by many stain separation-based methods for image normalization (Macenko et al. 2009, Khan et al. 2014) or color deconvolution (Ruifrok et al. 2001, Rabinovich et al. 2003). It is crucial to underline that, thanks to the exploitation of the a_1 and r_1 coordinates, the Φ-Norm normalization output is independent on the presence of different background white space in terms of both area and color expression. Indeed, coordinates a1 and r1 represent the images background colors, which are normalized by the proposed approach.

3) The coordinates (G_P,S_P) of each pixel projection P of the analyzed image can be expressed in terms of the vector A exploiting **Eq. 30** and **Eq. 31**, as shown in **Figure 30** (green formula). By taking advantage also of **Eq. 32**, the algorithm is able to retrieve the value of the weights a,b and c characterizing each pixel projection.

4) **Eq. 30** and **Eq. 31** are exploited again to map the coordinates P of points in the analyzed phasor plane to its transformed coordinates P' in the reference phasor space (**Figure 30C**): as shown in **Figure 30** in orange, the coordinates ($G_{P'}$,$S_{P'}$) are obtained by multiplying the weights a,b and c for the ordered components of the reference vector R.

5) Each pixel of the analyzed image is associated to the 8-bit RGB color whose coordinates on the phasor plot (**Figure 30D**) are nearest to P', in terms of Euclidean distance. In order to preserve the information and limit the degeneration effect, the retrieved color is mapped in the HSV space, and its Hue and Saturation parameters P'_{Hue} and $P'_{Saturation}$ are associated to the pixel.

6) The Value parameter is retrieved directly from the analyzed image: the RGB color referring to the pixel is transformed in the HSV space and the corresponding Value parameter $pixel_{Value}$ is exploited.

7) The triplet [P'_{Hue}, $P'_{Saturation}$, $pixel_{Value}$] associated to the pixel is mapped from the HSV space to the RGB space (**Figure 30E**) and the new vector [P''_{Red}, P''_{Green}, P''_{Blue}] represents the final stain normalized intensities of the red, green and blue channels of the pixel.

Once each pixel of the analyzed image has been processed, Φ-Norm provides the corresponding stain normalized image, in a 8-bit RGB format.

6.2 Staining normalization of H&E patches

In order to provide a proof of concept application of the proposed method and demonstrate its effectiveness in reducing the color difference in tissue images acquired in multiple laboratories, Φ-Norm has been applied to H&E-stained human liver tissue patches (1126x938 pixels). The colour content of the analysed images (an example is reported in **Figure 31B**), available in the repository of the University of Leeds (ULVPR), has been

normalized by exploiting a reference, which is an H&E-stained murine hepatocarcinoma section (**Figure 31A**), acquired at DKFZ (Heidelberg, Germany).

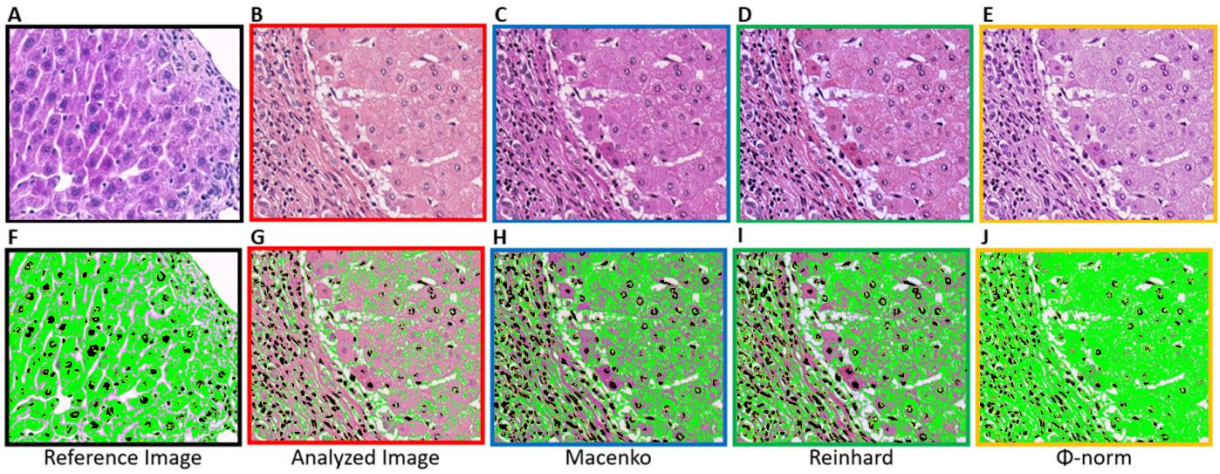

Figure 31: Staining normalization of an exemplary H&E patch *performed by exploiting the Macenko (**C**), Reinhard (**D**) and the proposed method (**E**). Panel **A:** region of interest of a murine liver tissue section, acquired at DKFZ and exploited as reference. Panel **B:** analyzed image related to an H&E stained human liver tissue section available in the ULVPR. Panels **F-J** report the results of a naive segmentation protocol based on color thresholding and applied on the images of panels A-E to identify the nuclei (highlighted in black) and the cytoplasm (in green) present in the tissue. Field of view: 248x206 µm^2 (1126x938 pixel^2).*

The staining normalization task has been performed by exploiting the proposed method (**Figure 31E**), the Reinhard approach (Reinhard et al**.** 2001**, Figure 31D**) and the Macenko protocol (Macenko et al. 2009, **Figure 31C**). From a qualitative perspective, the results achieved by the three different algorithms are compatible and affected by few differences. In particular, the normalized image retrieved by Macenko is slightly more reddish and saturated, while Reinhard and Φ-Norm show a more similar outcome. Compared to the reference image, the normalized background pixels obtained by Reinhard are darker and bluish, while Macenko algorithm and Φ-Norm show a brighter and more resembling background. In order to provide a more quantitative analysis, a naive segmentation of two biological structures, nuclei (**Figure 31F-J**, in black) and cytoplasmatic regions (**Figure 31F-J**, in green) has been obtained by exploiting a simple color thresholding protocol. In particular, pixels characterized by color intensities lower than [119, 91, 211] and [230, 244, 187] in the RGB colorspace have been associated to nuclei and cytoplasm, respectively.

These thresholds have been selected to reach the best accuracy in the Reference image and compare the results obtained by exploiting the same protocol on the analyzed image before and after the exploitation of different staining normalization algorithms. Concerning the nuclei segmentation, a reliable result is achieved also without taking advantage of staining normalization (**Figure 31G**), while the Φ-Norm color expression of the cytoplasmatic areas is the most similar to the Reference image: this is clearly showed in **Figure 31J**, where the segmentation accuracy of the cytoplasmatic regions is higher, compared to the results obtained by applying Macenko or Reinhard protocols (**Figure 31H** and **Figure 31I**, respectively). In particular, the Intersection over Union (IoU) parameter related to the cytoplasmatic region is 91.1% in the Φ-Norm-normalized image, compared to the 62.5% and 65.8% of the other investigated approaches (Macenko and Reinhard, respectively). Furthermore, the histograms of the pixel color content (**Figure 32**) have been computed for the reference (**Figure 31A**) and the normalized images obtained through the three reported methods (**Figure 31 C-E**). The three distributions related to the red channel (**Figure 32 A,D,G**) are very similar to the reference (in gray) and show a single peak around 200, while the histogram obtained through the Macenko method is also characterized by a second peak around the value 235. In the green channel (**Figure 32 B,E,H**), the histogram related to the Reinhard method is almost superimposed to the reference: they are characterized by the same peak position, around the value of 113, and by a similar width. The Macenko distribution is narrower, even if its peak is in the same position, around 113.

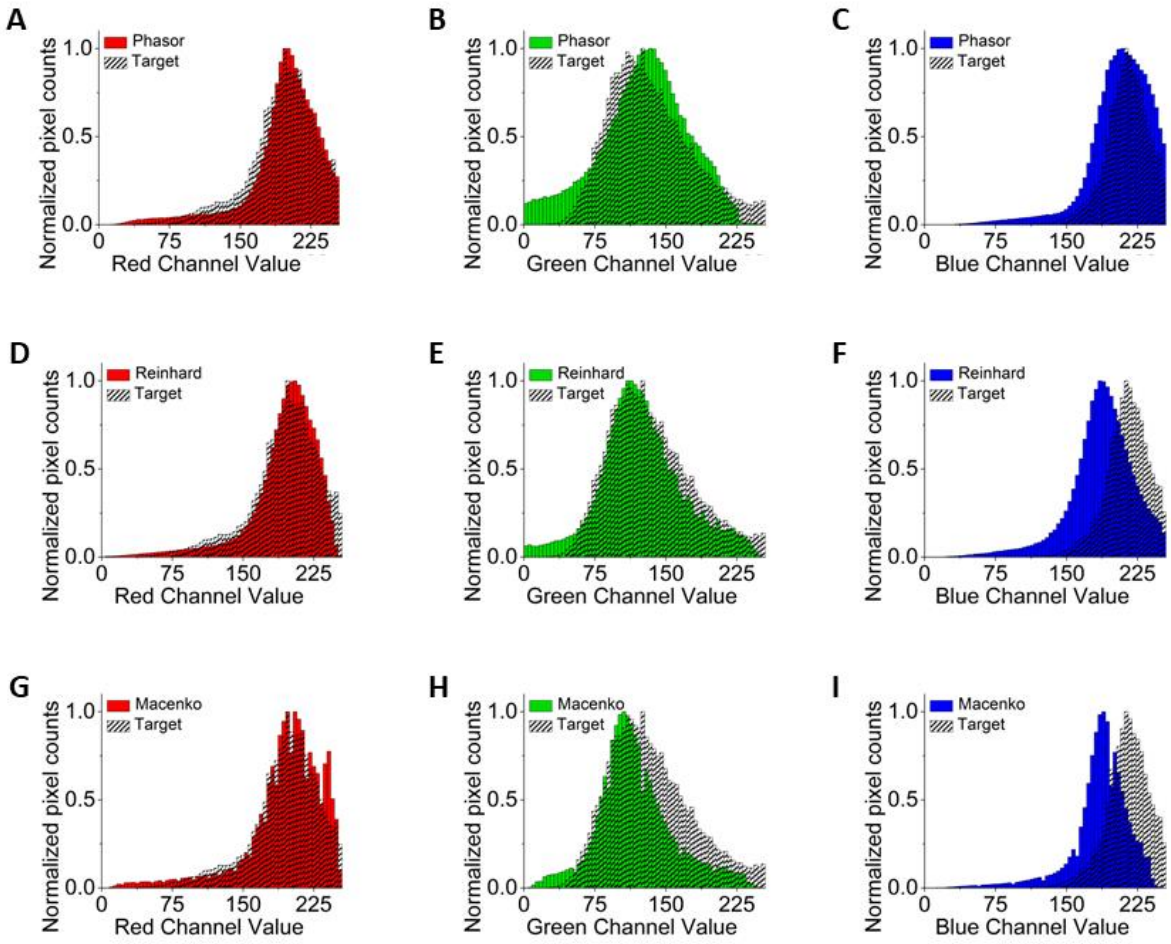

Figure 32: Normalized pixels intensity distributions of reference and normalized RGB images for each color channel. *Panels* ***A, B, C*** *show the pixels intensity distributions of the red, green and blue channels after Φ-Norm stain normalization (****Figure 31E****), while panels* ***D, E, F*** *refer to the normalized image obtained by exploiting Reinhard approach (****Figure 31D****). Panels* ***G, H, I*** *are the corresponding histograms of the red, green and blue channels related to the Macenko normalized image (****Figure 31C****). In each panel, the histograms of the red, green and blue channels related to the reference image (****Figure 31A****) are reported in black.*

Compared to the reference image, the green channel histogram related to Φ-Norm is the most different: even if the distribution width is comparable to the reference, its peak is shifted toward higher intensity values, around 131. However, in the blue channel (**Figure 32 C,F,I**), the scenario is inverted: only the peak position related to the Φ-Norm method (around 200) is comparable with the reference histogram, while Reinhard's and Macenko's algorithms show a single peak around 180.

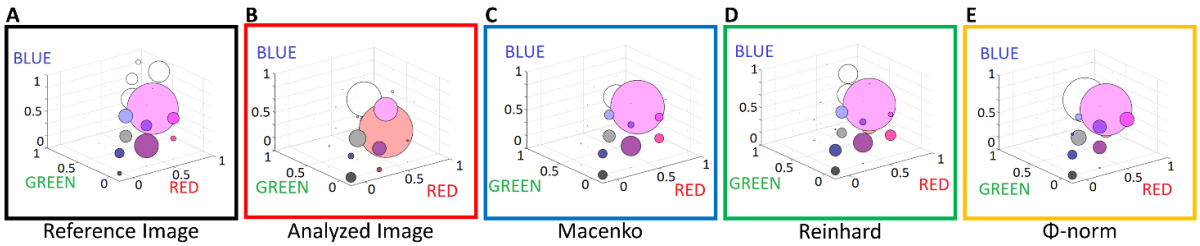

Figure 33: 3D pixel intensity histograms in the RGB colour-space. *The red, green and blue channels have been normalized in the interval [0 1] and separated in four equally spaced intervals, obtaining 64 groups represented by bubbles, color-coded by the mean color in the considered 3D interval. The volume of the bubbles is linearly proportional to the number of pixels in the clusters. Panel* ***A*** *refers to the distribution of the reference image (****Figure 31A****), while panel* ***B*** *corresponds to the analyzed image (****Figure 31B****). Panels* ***C, D ,E*** *represent the 3D color distributions of the images normalized by means of the Macenko (****Figure 31C****), Reinhard (****Figure 31D****) and the proposed method (****Figure 31E****), respectively.*

A different quantitative characterization of the methods has been obtained through a 3D visualization of the color distributions in the RGB space (**Figure 33**). The red, green and blue channels have been normalized in the interval [0 1] and separated in four equally spaced intervals ([0 0.25], [0.25 0.50], [0.50 0.75] and [0.75 1.00]), obtaining 64 groups which are represented by bubbles of different dimensions. The bubbles diameter is linearly proportional to the number of pixels in the corresponding group: a bigger bubble represents a cluster containing more pixels. The distributions (**Figure 33A and Figure 33B**) associated to the reference (**Figure 33A**) and the analysed images (**Figure 31B**) are very different. In both images, most of the pixels belong to few groups. The cluster with coordinates [0.63 0.38 0.63] contains the 52.0% of the pixels for the reference image, while only the 10.4% for the analysed image. On the other hand, the cluster characterized by the coordinates [0.63 0.38 0.38], where the intensity of the blue channel is lower, contains the 53.8% and the 3.4% of the pixels for the analysed and reference images, respectively. Since most of the pixels represent the cells cytoplasm in both images, the highlighted shift in the 3D distributions is caused by a difference in the Eosin color. This is also demonstrated by the qualitative observation of the patches: the colour corresponding to the Eosin staining in the reference image is identified by a higher contribute of the blue channel, while the analysed image appears reddish, due to the lack of the blue colour contribute. Regarding the Haematoxylin

staining, the differences between the reference and analysed images are more difficult to be visualized: from a qualitative observation, the nuclei of the analysed image appear slightly darker. However, this discrepancy is highlighted in the 3D distributions (**Figure 33A, Figure 33B**): the bubble of coordinates [0.13 0.13 0.13], which is associated to the dark blue and black colours, groups the 0.4% of reference image pixels and the 1.9% of the analysed image pixels. On the other hand, considering the bubble of coordinates [0.13 0.13 0.38], which identifies a lighter blue colour, the scenario is inverted: this cluster collects the 1.8% and the 0.7% of the pixels for the reference and analysed images, respectively. The 3D distributions of the normalized images (**Figure 33C-E**) are also suitable to demonstrate the effectiveness of the exploited staining normalization methods. A brief summary of the most crucial data is reported in **Table 4**.

Bubbles RGB coordinates	Reference image pixels [%]	Analyzed image pixels [%]	Macenko image pixels [%]	Reinhard image pixels [%]	Φ-norm pixels [%]
[0.63 0.38 0.63]	52.0%	10.4%	62.5%	56.2%	46.0%
[0.63 0.38 0.38]	3.4%	53.8%	2.7%	5.8%	3.6%
[0.13 0.13 0.13]	0.4%	1.9%	1.8%	1.4%	1.2%
[0.13 0.13 0.38]	1.8%	0.7%	1.7%	2.9%	1.4%

Table 4: percentage of pixels associated to four different clusters *(whose coordinates are reported in the table) of the 3D distributions reported in* ***Figure 43****.*

Compared to the reference image, the distributions related to the three staining normalization methods are very similar. In particular, as for the reference image, the cluster of coordinates [0.63 0.38 0.63], which represents the cytoplasm, is the biggest and only few pixels (<6%) belong to the [0.63 0.38 0.38] group. Moreover, for each staining normalization method, the cluster corresponding to the dark blue and black color is smaller than the group characterized by the [0.13 0.13 0.38] coordinates, representing the nuclei. It demonstrates that the colors of the analyzed image are successfully normalized by exploiting any of the three staining methods, strictly preserving the 3D distribution of the reference image colors.

6.3 Staining normalization of PicroSirius Red-stained images

In order to prove the robustness and versatility of the proposed method in performing the color normalization task for different staining protocols, the Φ-Norm, Reinhard and Macenko approaches have been applied to PicroSirius Red (SR)-stained whole slide images of murine liver tissue sections acquired at DKFZ (Heidelberg). The exemplary image in **Figure 34B** has been selected for the low quality of the SR stain, while the color normalization results are reported in **Figure 34C (**Macenko**), Figure 34D (**Reinhard**), Figure 34E (**Φ-Norm).

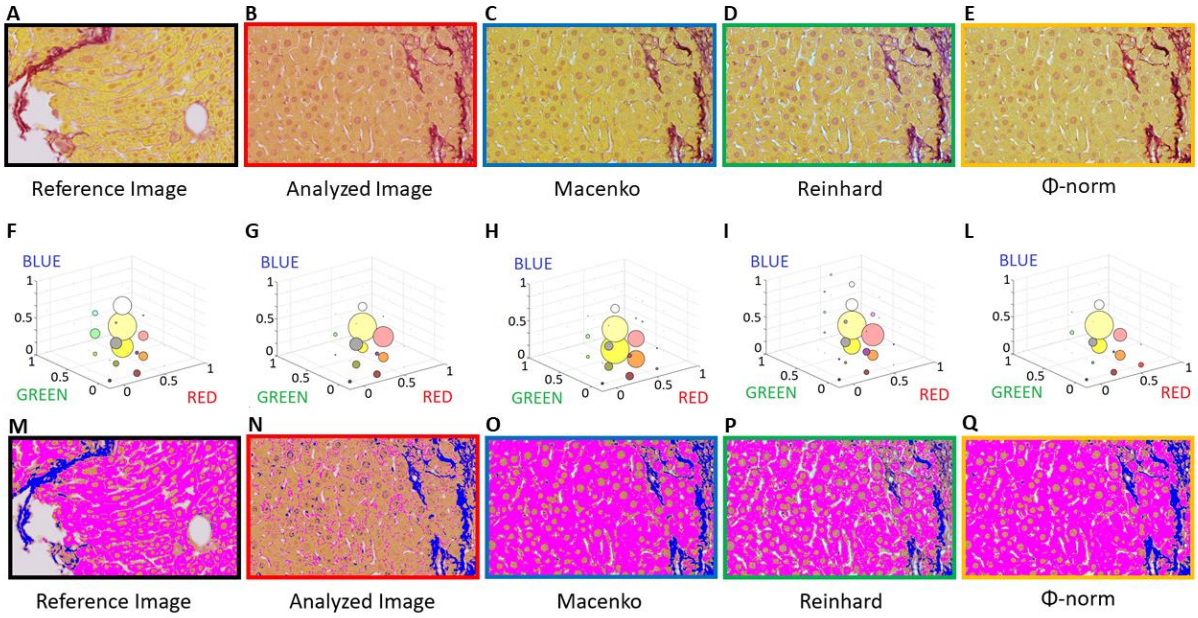

Figure 34: Stain normalization of an exemplary SR patch *performed by Macenko (**C**), Reinhard (**D**) and Φ-Norm (**E**). Panel **A:** region of interest of a murine liver tissue section exploited as reference. Panel **B:** analyzed image characterized by a low quality staining in terms of color content with respect to the reference. Panels **F-L**: 3D pixel intensity histograms in the RGB colour-space. Panel **F** refers to the distribution of the reference image (**Figure 34A**), while panel **G** corresponds to the analysed image (**Figure 34B**). Panels **H,I,L** represent the 3D distributions of the images normalized by means of Macenko (**Figure 44C**), Reinhard (**Figure 34D**) and Φ-Norm (**Figure 34E**). Panels **M-Q** report the results of a naive segmentation protocol based on color thresholding and applied on the images of panels A-E to identify the collagenous areas (highlighted in blue) and the cytoplasm (in magenta) present in the tissue. The images in panels **A-E** and **M-Q** are characterized by the same field of view: 156x95 μm^2 (708x432 pixel^2).*

The qualitative comparison between the reference (**Figure 34A)** and the analyzed images show that the collagen is characterized by a similar red color, while the cells cytoplasm suffers a huge discrepancy in terms of color content, as it is identified in yellow (reference) or orange (analyzed image). Moreover, the nuclei in the analyzed image are slightly darker than in the reference. These differences are also reflected in the 3D histogram visualization of the RGB colour-space. In the analyzed image distribution (**Figure 34F**), only the 6.1% of the pixels are grouped in the cluster of coordinates [0.63 0.63 0.13], which corresponds to the orange color, while it collects the 21.8% of the reference image pixels (**Figure 34G**). The cluster centered in [0.63 0.63 0.38], related to the light orange color, groups almost the same number of pixels (37.3% and 44.5% for the reference and analyzed image, respectively), while the light red cluster in [0.63 0.38 0.38] is more dense in the analyzed image distribution, collecting the 22.9% of the pixels. From a qualitative perspective, the Reinhard method appears less effective in performing the normalization task: the collagen is characterized by a magenta color (**Figure 34D**) instead of red, which means that the blue channel intensity is overestimated. However, the cytoplasmatic region is successfully normalized to yellow and the nuclei orange color has been preserved, as expected. Φ-Norm (**Figure 34E**) and Macenko (**Figure 34C**) methods, which are stain vectors-based, perform a more compatible color transformation, mapping the cytoplasmatic region from orange to yellow and preserving, at the same time, the color content of both nuclei and collagen areas. Discrepancies in the color normalization procedures are easily highlighted by exploiting 3D histograms: the magenta cluster centered in [0.63 0.38 0.38] groups more pixels in the Reinhard distribution (23.6% - **Figure 34I**) than the other staining normalization approaches (11.3% and 10.8% for Macenko and Φ-Norm, **Figure 34H** and **Figure 34L**, respectively) since Reinhard method identifies collagen with a magenta color, instead of red, as correctly performed by Macenko and Φ-Norm. Similarly to the H&E-stained sections, a naive segmentation, based on color thresholding, has been applied to SR-stained images to identify collagen and cytoplasmatic areas present in the tissue (**Figure 34 M-Q**). In this case, pixels characterized by color intensities lower than [255, 114, 104] and [245, 241, 100]

in the RGB colorspace have been associated to collagen and cytoplasm, respectively. Concerning the cytoplasm segmentation, no discrepancies have been highlighted among the results obtained through the different staining normalization methods (**Figure 34 O-Q**), while the collagen segmentation after the image processing involving the Reinhard algorithm (**Figure 34P**) is characterized by a really low accuracy, if compared to the other protocols. This is quantitatively demonstrated by the IoU related to the collagenous tissue present in the image since it is high for the images which are normalized by exploiting Macenko and Φ-Norm methods (94.2% and 94.8%, respectively) and lower for Reinhard (71.0%).

6.4 Φ-Norm application to improve segmentation protocols accuracy

The aim of this analysis is to demonstrate that, compared to the state of the art staining normalization protocols, Φ-Norm is a useful preprocessing tool to improve the performance of Machine Learning-based semantic segmentation methods. Indeed, these approaches consider that the same biological structure is characterized by the same color content in any biopsy. Unfortunately, this is not true in experimental conditions, due to discrepancies in the staining procedures performed by different laboratories or to different image acquisition scanners.

During this analysis, performed for both H&E and SR staining, different biological structures (nuclei, cytoplasmatic area and collagen) have been manually segmented by an expert in a reference image, acquired in the DKFZ laboratories, and in other 20 images, collected in the ULVPR, in the TGCA or in the DKFZ repositories (**Figure 35)**.

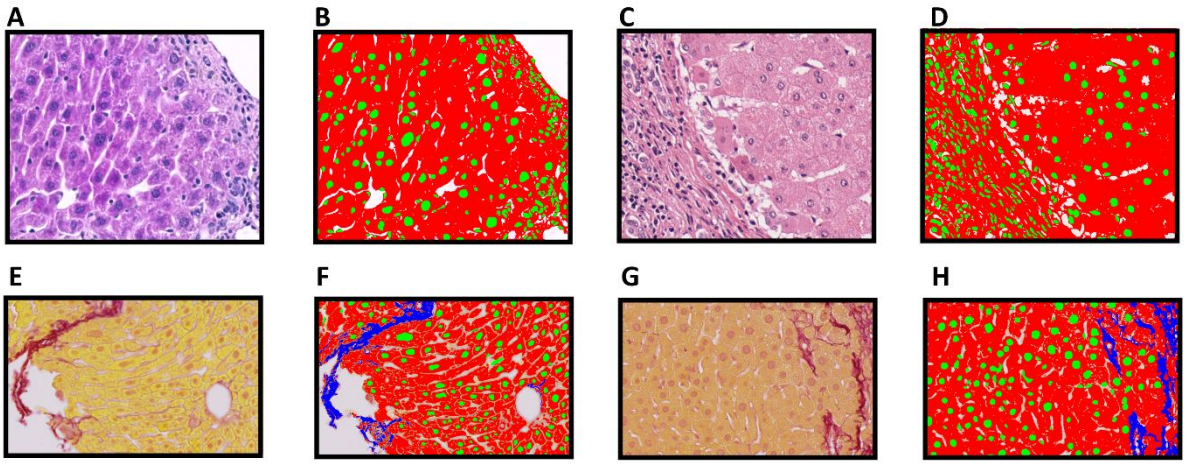

Figure 35: Manual segmentation of biological structures performed by an expert. ***A*** *and* ***E*** *are the H&E and SR-stained patches extracted from murine biopsies acquired at the DKFZ laboratories and exploited as reference images, while* ***B*** *and* ***F*** *are the corresponding outcomes of the manual segmentation performed by the expert. The pixels classified as nuclei are green-colored, while the regions recognized as cytoplasm and collagen are reported in red and blue, respectively. Macenko, Reinhard and Φ-Norm approaches have been exploited to analyze 20 H&E (248x206 μm^2) and 20 SR (156x95 μm^2) stained images. In particular,* ***C*** *and* ***G*** *are exemplary analyzed H&E and SR-stained ROIs, respectively.* ***D*** *and* ***H*** *are the corresponding outcomes of the manual segmentation performed by the expert, color-coded as* ***B*** *and* ***F****.*

Then, the 20 images have been staining-normalized by applying Reinhard, Macenko and Φ-Norm methods. The mean intensity values of each red, green and blue channels have been computed for each structure, manually segmented by the expert, both in the reference and in the color normalized images, (**Figure 36** and **Figure 37** for the H&E and SR-stained images. respectively).

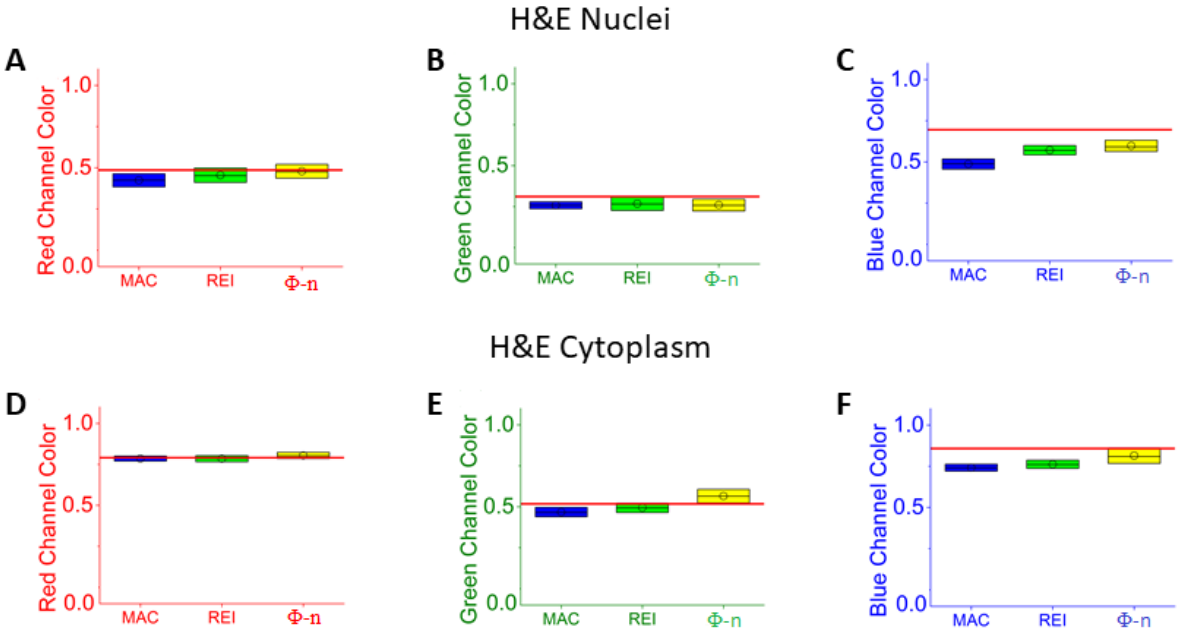

Figure 36: Boxplots reporting the mean, median and standard deviation computed for the nuclei (A, B, C) and the cells cytoplasm (D, E, F) clusters, manually segmented by an expert in 20 H&E images after the normalization procedure. *In particular, panels **A, D** are related to the red channel color component, while (**B, E**) and (**C, F**) refer to the green and blue channels, respectively. In each graph, the reference image mean intensity of the channel is represented by a red line, The median and the mean values are represented by a circle and a horizontal line inside the boxes, whose edges represent the upper and lower limits of the standard deviation.*

Concerning the H&E-stained images, the red channel (**Figure 36D**) distributions of the three methods are compatible with the reference value (red line in **Figure 36**) for the cells cytoplasm. On the other hand, for nuclei (**Figure 36A**), the Macenko method slightly underestimates the mean intensity, while the Reinhard and Φ-Norm are compatible with the reference value. In the green channel, the outcomes of the Reinhard and Φ-Norm approaches agree with the reference image better than the Macenko method, both for nuclei (**Figure 36B**) and cells cytoplasm (**Figure 36E**). Finally, regarding the blue channel, the trends for the nuclei (**Figure 36C**) and the cells cytoplasm (**Figure 36F**) are very similar. In particular, in both cases, Φ-Norm shows the best agreement with the reference, while Macenko is characterized by the least compatible results.

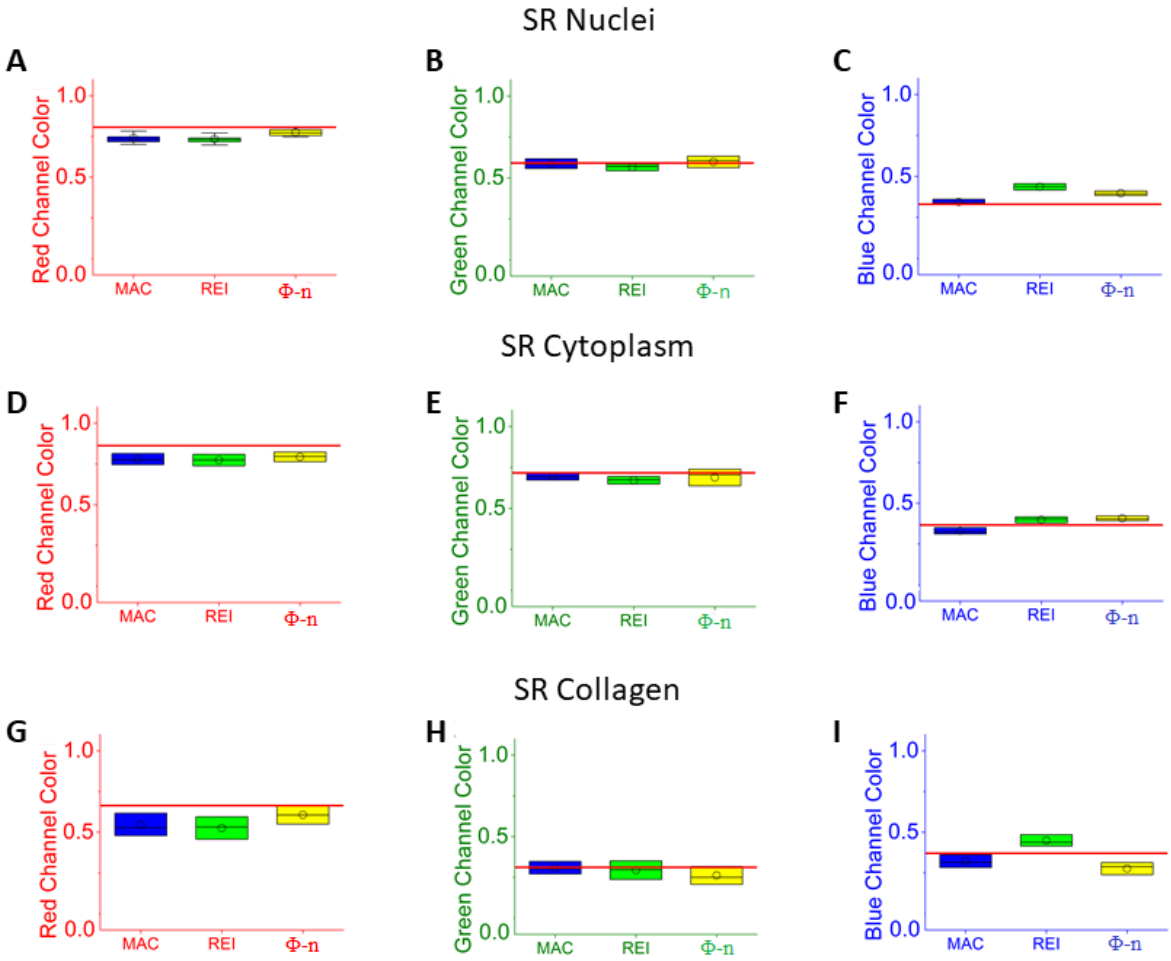

Figure 37: Boxplots reporting the mean, median and standard deviation computed for the nuclei (A, B, C), the cells cytoplasm (D, E, F) and the collagen (G, H, I) clusters manually segmented by an expert in 20 SR images after the normalization procedure. *In particular, panels (**A, D, G**) are related to the red channel color component, while (**B, E, H**) and (**C, F, I**) refer to the green and blue channels, respectively. In each graph, the reference image mean intensity of the channel is represented by a red line. The median and the mean values are represented by a circle and a horizontal line inside the boxes, whose edges represent the upper and lower limits of the standard deviation.*

By taking into account the SR results (**Figure 37**), Φ-Norm agrees with the reference image red color content better than the other methods. In particular, **Figure 37A** and **Figure 37G** highlight that Φ-Norm is totally compatible with the reference value, while Reinhard and Macenko slightly underestimate the red intensity. The best performance in the green channel is accomplished by the Macenko and the Φ-Norm approaches, whose distributions are compatible with the reference image mean intensity in every segmented biological structure (**Figures 37B, E, H**). In the blue channel, the mean color intensity of the cells cytoplasm (**Figure 37F**) in the reference image is similar to the outcomes of the three normalization methods. However, the intensity in the nuclei (**Figure 37C**) are slightly overestimated by all

the normalization procedures with Macenko showing the best compatibility with the reference. Concerning the collagen regions, the blue channel (**Figure 37I**) is slightly underestimated by Macenko and Φ-Norm, while it is overestimated by Reinhard.

Since the color characterizing a biological structure is the result of the different superposition of the red, green and blue contributes, a metric able to simultaneously consider discrepancies affecting the three (R, G, B) color contents has been conceived and named Color Mean Distance (CMD). Indeed, the (R, G, B) components of each pixel can be represented in a 3D color space, where different biological structures with similar colors are grouped in clusters. The centroid coordinates correspond to the mean color of the clusters and therefore of the biological structure. By considering two different clusters lying in the same N-dimensional colour-space (N=3 for RGB) and characterized by the centroids [x_1, x_2, …, x_N] and [y_1, y_2, …, y_N] respectively, their CMD value is computed as their Euclidean distance:

$$CMD = \sqrt[2]{\sum_{i=1}^{N} (x_i - y_i)^2} \qquad (34)$$

If the two clusters correspond to the same biological structure characterized by the same color in different images the CMD value will be zero. However, the centroids distance will increase if the color content discrepancy between the two clusters grows and the CMD will be characterized by a higher value. In our case, the CMD between the centroids of the biological structures after the stain normalization protocols and those related to the reference has been computed for 20 H&E and 20 SR stained patches. A summary of the computed CMD values are reported in **Table 5**.

CMD value H&E images	Macenko	Reinhard	Φ-norm
Nuclei	0.24 ± 0.03	0.16 ± 0.01	0.13 ± 0.01
Cytoplasm + Collagen	0.13 ± 0.02	0.11 ± 0.01	0.08 ± 0.01
CMD value SR images	**Macenko**	**Reinhard**	**Φ-norm**
Nuclei	0.08 ± 0.01	0.14 ± 0.01	0.08 ± 0.01
Cytoplasm	0.09 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
Collagen	0.15 ± 0.02	0.18 ± 0.01	0.13 ± 0.02

Table 5: Mean CMD values referred to the biological structures manually segmented in 20 H&E and 20 SR stained patches. *The CMD has been computed between the centroids representing the considered biological structure for the stain normalized and the reference images.*

Regarding the H&E stained tissue sections, Φ-Norm is characterized by the best results among the considered staining normalization methods. In particular, compared to the Macenko approach, the color content discrepancy from the reference values is reduced by ~50% for the nuclei and ~40% for the cells cytoplasm. Furthermore, the three methods are able to better resemble the cytoplasm color of the reference image, while the mean CMD obtained in the nuclei is higher for each approach.

Concerning the SR images, the Macenko and Φ-Norm approaches are characterized by similar performances in terms of CMD. Compared to Reinhard, the Φ-Norm color discrepancy from the reference image is reduced by ~40% for the nuclei and by ~25% for collagen, while the differences among the three approaches are not relevant for the cells cytoplasm.

6.5 Staining normalization of Whole Slide Images

Finally, the efficacy of Φ-Norm in stain normalization of entire whole slide images is demonstrated in **Figure 38** and **Figure 39** for two different stainings: PicroSirius Red and H&E, respectively. In the first case, Φ-Norm has been applied to uniform the color expression of the same whole slide section, acquired at the DKFZ laboratories. In particular,

a single patch of 512x512 pixels extracted from a region near the sample boundaries has been exploited as reference to analyze the entire slide. In the second scenario an entire H&E-stained whole slide section of the DKFZ laboratories has been exploited as reference to modify the color palette of a sample acquired at the Humanitas Research Hospital (Prof. Luca Di Tommaso, Milan – Italy). In both cases, the whole analyzed slides have not been divided in patches during the normalization procedure, as Φ-Norm acts pixelwise in the phasor space domain.

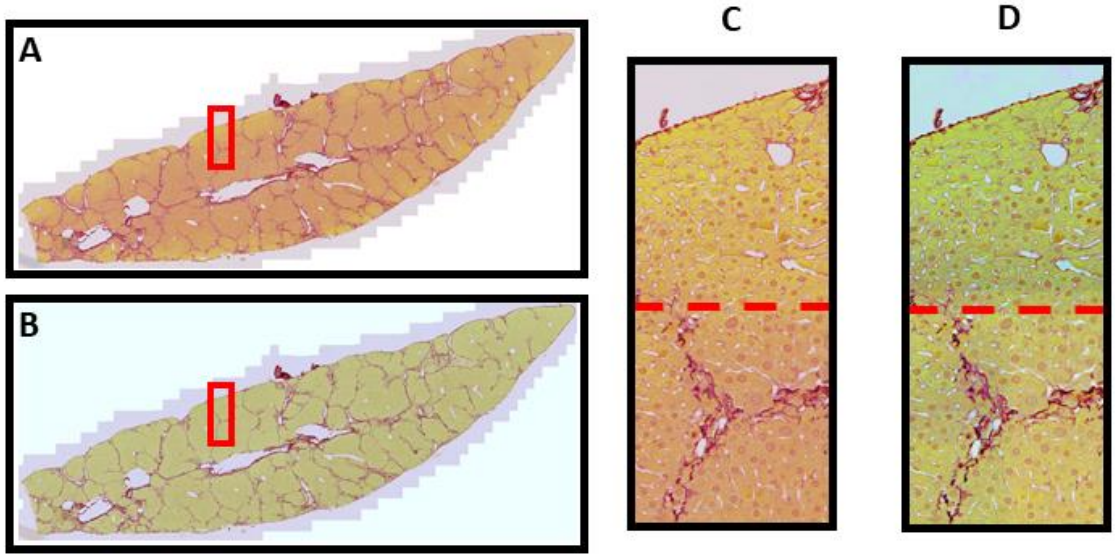

***Figure 38**: **Application of the Φ-Norm method to an exemplary SR-stained whole slide image.** Panel **A** shows the analyzed image (4.0x1.7 mm^2), while **B** reports the results of the Φ-Norm method. Panel C shows the region of interest (156x265 μm^2) highlighted by a red rectangle in panel A, while panel D shows the same patch extracted from the Φ-Norm normalized image in panel B. The horizontal red dashed line **in C** and **D** separates two analyzed regions.*

In particular, concerning the PicroSirius Red staining, the ROI extracted from the analyzed image (**Figure 38C**) and highlighted by a red rectangle in the biopsy (**Figure 38A**) reveals clear discrepancies in the color content of the cells cytoplasm, which is orange from the center to the lower part of the image, while it is yellow in the upper part. As a consequence, a common color-based threshold is not effective in this case to perform an accurate semantic segmentation of the biological structures. The stain normalization results obtained through Φ-Norm are shown in **Figure 38B.** After the normalization procedure, the yellow color related to the cytoplasmatic region is uniform through the entire sample. These

considerations are also qualitatively proved in the zoomed region reported in **Figure 38D**, where the discrepancies of color content between the upper and the lower part of the image, separated by a red dashed line, are clearly reduced. In order to provide a quantitative description of the obtained color normalization, the cells cytoplasm has been manually segmented by an expert in **Figure 38C** and **Figure 38D**. The mean intensities of the red, green and blue channels have been computed for the lower and upper areas of the two considered patches, obtaining the values [0.79 0.59 0.42] and [0.84 0.71 0.33] for the non-normalized ROI, while [0.82 0.78 0.26] and [0.86 0.83 0.28] after Φ-Norm. In order to better quantify the color normalization provided by the application of Φ-Norm, the CMD parameter between the mean color values of the upper and lower part of both patches has been computed. In particular, CMD = 0.16 for the non-normalized ROI (**Figure 38C**), while CMD = 0.07 after the application of Φ-Norm (**Figure 38D**), demonstrating that the color discrepancy of the cells cytoplasm in the upper and lower region of the patch has been reduced by more than 55%.

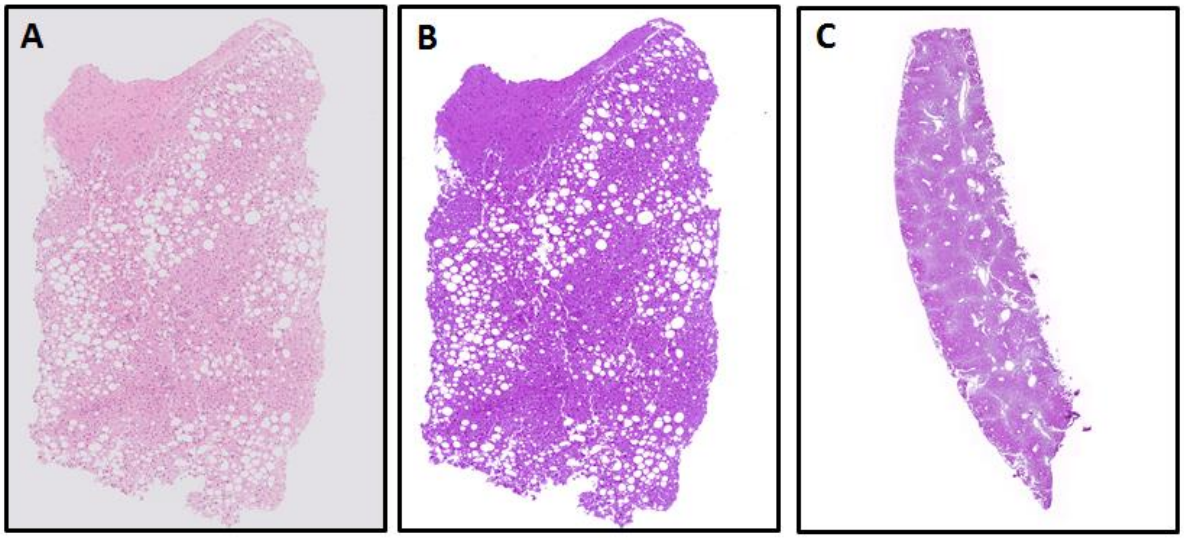

***Figure 39**: **Application of the Φ-Norm method to an exemplary H&E whole slide image.** Panel **A** shows the analyzed image (1.3x1.8 mm^2), while **B** reports the results of the Φ-Norm method. Panel **C** is the whole slide sample which has been exploited as reference (1.7x4.1 mm^2).*

Regarding the Φ-Norm application to the H&E staining, a qualitative assessment of the normalization results provided by the proposed method (**Figure 39B**) confirms that both the

color expressions of nuclei and cytoplasm (dark blue and violet) are similar to the reference image (**Figure 39C**). Similarly to the PicroSirius Red experimental scenario, the CMD parameter has been computed between the reference and the normalized image for the cytoplasm (CMD = 0.06). In this case, the color discrepancy has been reduced by more than 80%.

The analysis performed demonstrated that Φ-Norm is the most suitable method among the investigated approaches to be coupled with AI-based techniques. Indeed, it showed the best CMD values for both H&E and SR stained biopsies. Moreover, Φ-Norm can effectively improve the diagnosis clinical routine since it is a fully automatic pipeline and can be applied to whole-slide sections by pathologists with standard computational knowledge. In this context, the computational time needed to apply Φ-Norm for a single whole-slide image is lower than 5 minutes, if a standard computer is exploited. As a consequence, more than 300 samples could be normalized and then evaluated each day by the expert. Furthermore, the application of Φ-Norm improves the accuracy of the expert analysis since color artifacts due to the experimental staining procedures are strongly reduced.

7. SuperμMAPPS: a novel tool to investigate collagen micro-architecture

The aim of this chapter is to demonstrate that the SuperμMAPPS algorithm is a reliable and robust tool to provide accurate analysis of the collagen micro-architecture, without the need of image preprocessing protocols and computational expertise since it is a fully automated approach. In particular, a detailed description of the SuperμMAPPS architecture is provided in the following. Then, in order to validate the proposed approach and provide its complete characterization, SuperμMAPPS has been applied to synthetic P-SHG images by simulating different acquisition conditions. During this analysis, the method accuracy has been evaluated in terms of Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |V_{i(ret)} - V_{i(true)}| \qquad (35)$$

where N is the number of measurements, $V_{i(ret)}$ and $V_{i(true)}$ are the retrieved and the expected parameter values, respectively. The handling of the noise effect performed by SuperμMAPPS on the phasor space has been compared with two different denoising procedures: 1) NOISE2VOID (Krull et al. 2020), which is a neural network acting on the image plane and devoted to the mitigation of white Gaussian noise, and 2) an iterative median filter-based denoising method acting on the phasor (g,s) coordinate matrices G_{map} and S_{map}, (Cutrale et al. 2017; the elements of these matrices are the real and imaginary parts of the DFT applied to the P-SHG spectra, as computed through **Eq. 5**). Briefly, NOISE2VOID is a U-Net Neural Network-based method devoted to image denoising. Differently from many other neural network-based denoising approaches, NOISE2VOID is directly trained on the experimental data, without the need of clean target images: this property is crucial in the biomedical field since the acquisition of a huge amount of training data is often not possible. Due to its architecture, NOISE2VOID is not suitable to analyze structured noise since it violates the assumption of pixelwise independence.

SuperμMAPPS has been applied in simple experimental conditions to analyze the P-SHG images of murine tendons, whose collagen architecture is simple and well-defined. Furthermore, the effect of preprocessing denoising procedures on the SuperμMAPPS retrieving has been investigated: the same images have been processed by exploiting NOISE2VOID neural network (batch size 48x48, 100 epochs) to decrease the noise affecting each acquired signal and the results provided by SuperμMAPPS have been compared with the parameters extracted without the application of denoising methods. Moreover, images related to human tissue samples (breast carcinoma, skin tissue, thyroid), acquired by a different laboratory and stored in the PSHG-TISS public repository (Hristu et al. 2022), have been processed. In this case, SuperμMAPPS results have been compared with the outputs of standard fitting procedures. Finally, SuperμMAPPS has been applied in a real clinical context in order to demonstrate its capability in providing collagen-related features able to recognize samples characterized by early-stage breast carcinoma. In this case, patches of human breast biopsies characterized by three different tumor growth stages (Normal Terminal Ductal-Lobular Units – Normal TDLU, which is the healthy tissue, early stage and invasive tumorous stage) have been extracted from the PSHG-TISS and analyzed, demonstrating the capability of SuperμMAPPS to highlight early stage tumorous tissue, hardly detected by histopathologists during the semi-quantitative analysis of stained biopsies.

7.1 SuperμMAPPS algorithm

Method overview

SuperμMAPPS (SUPERvised Microscopic Multiparametric Analysis by Phasor projection of Polarization-dependent SHG) builds on the idea of i-φ-MaLe algorithm to analyze the P-SHG signal and provide quantitative parameters describing the collagen micro-architecture in biological tissues. Differently from the R_{app} spectra analyzed by i-φ-MaLe, characterized by hundreds of experimental points, the P-SHG signal is composed by 10-20 points. As a

consequence, considering that i-φ-MaLe provides optimal results by analyzing more than 50 different spectral windows composed by at least 10 experimental points (**Section 5.2** for further information), the application of the same statistical-based approach is not anymore possible, due to the lack of experimental data.

In this scenario, SuperμMAPPS has been devised to overcome this problem, providing reliable retrievals of parameters characterizing the collagen microarchitecture through the analysis of P-SHG spectra composed by only 6 points (**Section 7.3**).

Architecture

A schematic of the method is provided in **Figure 40**.

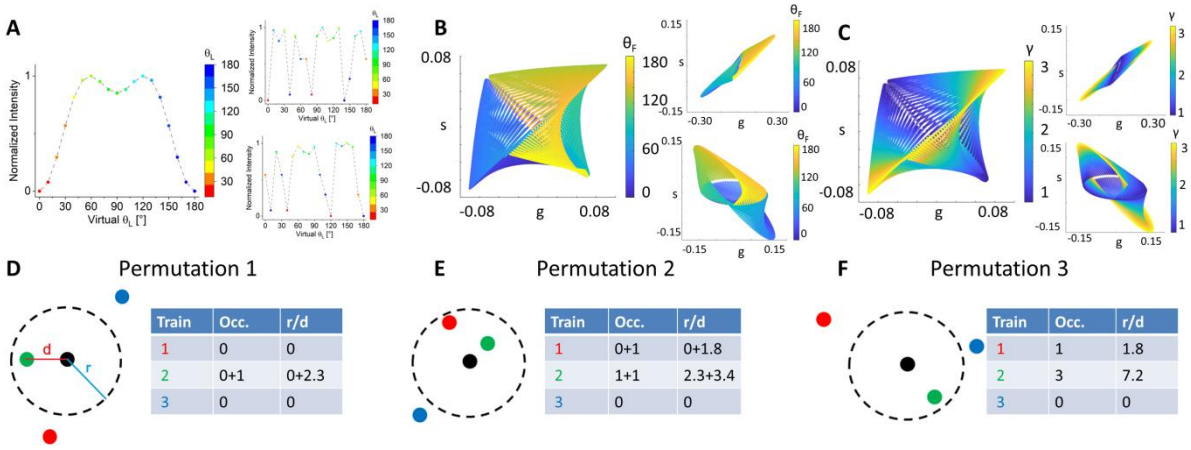

Figure 40. Schematic of the SuperμMAPPS method. *(**A**) A P-SHG spectrum and two related permutations are projected in different phasor plots (reported in **B**) by applying the DFT on the entire function. A supervised ML algorithm associates the θ_F and γ parameters values of the P-SHG spectra to the projected points in the phasor plane, shown here color-coded for θ_F in **B** and γ in **C**. This procedure is repeated for each P-SHG spectrum in the training dataset, obtaining the reported distributions. Panels (**D**)-(**F**) sketch the retrieving procedure for three different phasor plots originated by the three different permutations showed in A. The algorithm assigns to the unknown test point (in black) the θ_F and γ values corresponding to the trained point with the highest occurrence and distance scores, as already detailed in Section 5.1 (Figure 13).*

Similarly to i-φ-MaLe, SuperμMAPPS has been trained by means of a set of simulated functions. In this case, more than 35000 simulated P-SHG spectra have been obtained through **Eq. 24** while varying the θ_F parameter from 0° to 180° in steps of 1° and γ from 1.04 to 3.00 in steps of 0.01. Then, SuperμMAPPS computes the DFT of each entire P-SHG

spectrum (I_{PSHG}) of the training dataset in the interval [θ_{Ls} θ_{Lf}], obtaining on the phasor plot a distribution of points (**Figure 40B** and **Figure 40C**), whose coordinates are:

$$g\,(k) = \frac{\sum_{\theta_{Ls}}^{\theta_{Lf}} I_{PSHG}(\theta_L) cos(\omega k \theta_L)}{\sum_{\theta_{Ls}}^{\theta_{Lf}} I_{PSHG}(\theta_L)} \qquad s\,(k) = \frac{\sum_{\theta_{Ls}}^{\theta_{Lf}} I_{PSHG}(\theta_L) sin(\omega k \theta_L)}{\sum_{\theta_{Ls}}^{\theta_{Lf}} I_{PSHG}(\theta_L)} \qquad (36)$$

Each point on the phasor plot can be color-coded for the θ_F or γ values originating the P-SHG spectrum, as shown in **Figure 40B** and **Figure 40C**.

In this case, differently from i-φ-MaLe, multiple phasor plots are obtained by computing the DFT of different permutations of the original P-SHG spectrum. In every new phasor plot, each training dataset spectrum is projected after being modified by means of the same permutation function. As a consequence, if N experimental points characterize the analyzed P-SHG signal, a total of N! phasor plots can be obtained: this approach guarantees a sufficient number of phasor plots to apply the same statistical-based retrieving procedure exploited by i-φ-MaLe. Indeed, 10 experimental points, commonly acquired during standard P-SHG signal acquisition procedures, are able to originate more than 3.5 millions different phasor plots. Validation procedures highlighted that only few hundreds phasor plots are needed to obtain reliable results (**Figure 40** of **Section 7.2** for further information).

Similarly to i-φ-MaLe, SuperμMAPPS automatically takes into account the noise characterizing each experimental measurement by considering, during the retrieval phase, an uncertainty circle around the projected point with a radius inversely proportional to the SNR value (Cutrale et al. 2017). In particular:

1- A circle with radius r, centered around the test point (black dot in **Figures 39D-F**), is considered in each phasor space.
2- Each training dataset point inside the circle is associated to two different scores, the Occurrence score Occ and the Distance score D, computed through **Eq. 27**.

3- The algorithm associates to the test spectrum the single parameter couple $P_T=\{\theta_F, \gamma\}$ corresponding to the training point with the highest Occ score. In case of tie, the algorithm selects the values corresponding to the nearest point among those with the same occurrence. If a tie is reached also for D, the algorithm computes their mean value.

7.2 Method validation and analysis of synthetic images

SuperμMAPPS performances have been characterized by estimating the retrieval accuracy of both θ_F and γ parameters and by quantifying the computational time of the training and test phases (**Figure 41**). In particular, a total of 10000 simulated P-SHG spectra have been exploited as test set, with the θ_F and γ values randomly extracted from the same intervals used to create the training dataset, which is instead composed by more than 35000 simulated P-SHG spectra obtained by varying the θ_F parameter from 0° to 180° in steps of 1° and γ from 1.04 to 3.00 in steps of 0.01. Concerning the polarization angle θ_L, each training spectrum is composed by 18 points, simulated in the range [0° 180°] in steps of 10°.

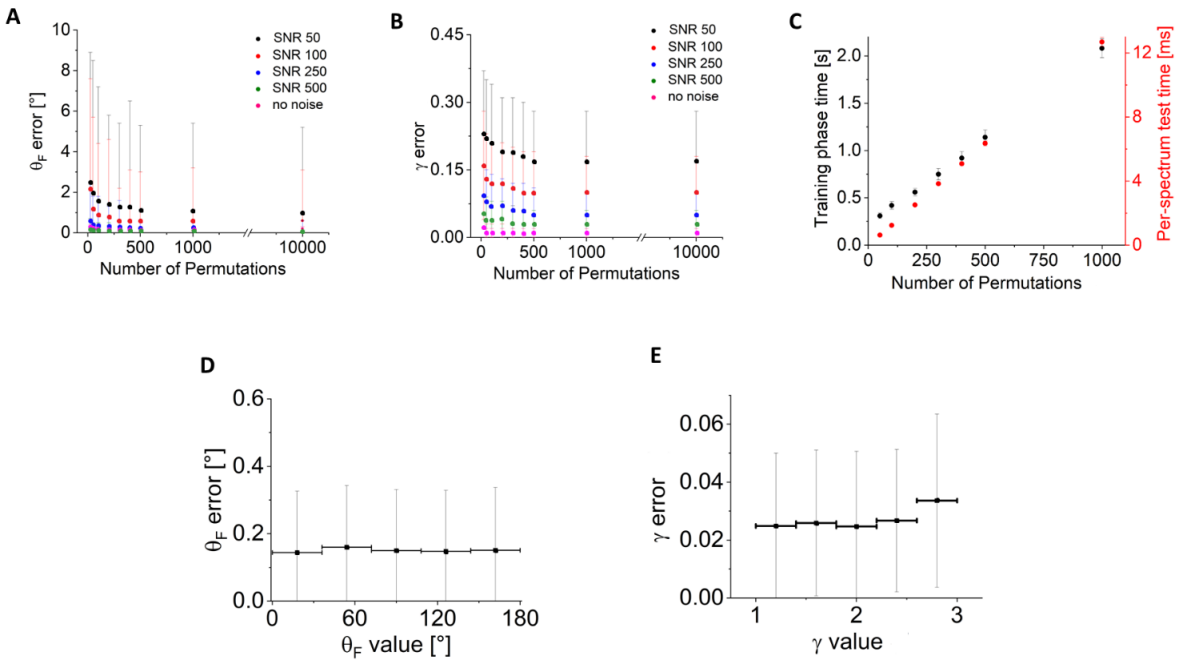

Figure 41: SuperμMAPPS hyperparameters characterization. *The method accuracy has been quantified in terms of MAE for both θ_F (**A**) and γ (**B**). An increasing noise (SNR = [50 500]) and a different number of permutations ([50 10000]) have been considered during the analysis. **C** reports the computational time characterizing the training (black dots) and the test (red dots) phases for an increasing number of permutations ([50 1000]). In **D** and **E** the MAE has been reported for different intervals of θ_F and γ values, respectively. θ_F MAE has*

been estimated in the range [0° 180°], divided in 5 steps of 36°, while γ MAE has been quantified in the range [1.04 3.00], divided in 5 steps of 0.4.

As shown in **Figure 41A** for θ_F and **Figure 41B** for γ, the retrieval accuracy has been estimated in terms of MAE (**Eq. 35**). The analysis has been performed by considering different algorithm configurations characterized by a varying number P of permutations, which is the only hyperparameter of SuperμMAPPS. For each configuration, the retrieval accuracy has been also investigated in dependence of an increasing noise affecting the test spectra. As expected, for a fixed number of permutations, both θ_F and γ MAEs increase if the analyzed spectra are affected by a higher noise. However, even if few permutations are considered (< 100), the mean MAE characterizing SuperμMAPPS is < 3° for θ_F and < 0.3 for γ. Moreover, for P > 250, the θ_F and γ accuracies are almost constant: 0.6° ± 2.0° and 0.10 ± 0.09 for SNR = 100, respectively. These considerations demonstrate that only 6 different values of θ_L have to be experimentally acquired for each P-SHG signal since, in this condition, a total of 6! = 720 permutations can be computed, allowing a reduction of the data acquisition time. Furthermore, the retrieval accuracy of θ_F and γ is uniform in the whole value intervals [0° 180°] and [1.04 3.00], as highlighted in **Figure 41D** and **Figure 41E** respectively. The computational time characterizing both the training and test phases (**Figure 41C**) scales linearly while increasing the number of permutations. For 500 permutations, where the retrieving errors for both θ_F and γ already reached a low value (0.6° ± 2.0° and 0.10 ± 0.09 for SNR = 100, respectively), a single P-SHG spectrum is analyzed in only ~ 6 ms.

Since experimental images are characterized by noise, a proper description of the SuperμMAPPS performances in these conditions has to be addressed. At the state of the art, many algorithms have been developed in order to reduce the noise contribute, as NOISE2VOID (Krull et al. 2020), which exploits neural networks, or an iterative median filter-based denoising method acting on the phasor (g,s) coordinate matrices G_{map} and S_{map} (Cutrale et al. 2017). Unfortunately, these denoising methods exploit the spatial information, assuming that contiguous pixels are characterized by similar values. Experimentally, this

assumption is not always true and can lead to the presence of artifacts. The following analysis is devoted to demonstrate that SuperμMAPPS takes into account the noise contribution directly in the phasor space, avoiding the artifacts caused by a priori considerations concerning the pixels spatial information. In order to verify this statement, two different synthetic images, affected by white Gaussian noise, have been analyzed.

The first P-SHG image series, composed of 50x50 pixels, represent a simple scenario where each pixel has been associated to the same P-SHG function (**Eq. 24**) composed of 18 equidistant points in the range [0° 180°] ($\Delta\theta_L$ = 10°), with θ_F = 60° and γ = 1.7. In this image, the spatial information regarding the pixels is not able to introduce artifacts since they contain the same P-SHG signal. At first, SuperμMAPPS has been applied with 500 permutations to retrieve the (g,s) coordinates related to the θ_F and γ variables without any denoising algorithm (**Figure 42**, in red). Then, the same quantification has been performed by applying NOISE2VOID (**Figure 42**, in green) or a 5 iterations median filter (Cutrale et al. 2017, **Figure 42** in blue) to the synthetic images before retrieving the (g,s) coordinates.

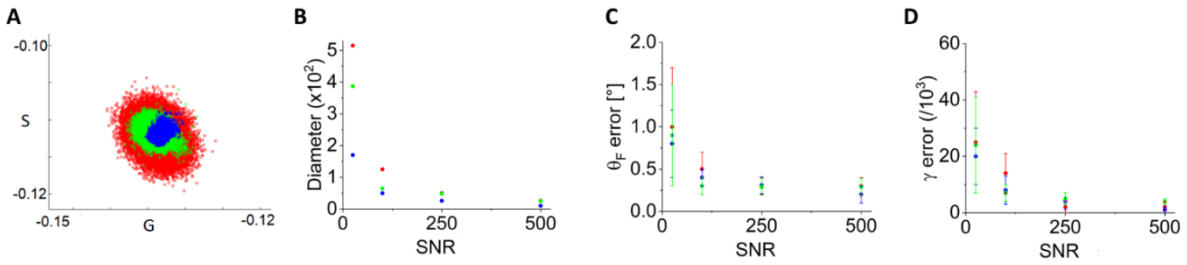

Figure 42: Results concerning the synthetic image characterized by the same P-SHG spectrum in each pixel. *Panel **A** shows the distribution of the P-SHG spectra projection (θ_F = 60° and γ = 1.7 in each pixel) for the entire image for SNR = 100 while SuperμMAPPS is applied (red), after the application of the NOISE2VOID algorithm on the image (green) and after the exploitation of the median-based filter method (blue). Panel **B** reports the mean diameter of the points distribution in dependence of the SNR value. **C** and **D** report the MAE corresponding to θ_F and γ, respectively, related to the (g,s) coordinates retrieved through SuperμMAPPS (red), through the DFT computation following the application of NOISE2VOID (green) and by applying the median-based filter method on the (g,s) coordinates obtained by applying the DFT (blue).*

For every algorithm configuration, the error associated to θ_F and γ grows for decreasing values of SNR, as theoretically expected. Generally, the application of denoising methods

lead to slight better results in terms of retrieval accuracy. Indeed, for SNR = 25, SuperμMAPPS method is characterized by a MAE of 1.0 ± 0.7 for θ_F (**Figure 42C**), while the application of NOISE2VOID or the exploitation of the median filter lead to a slight decrease of MAE (0.9 ± 0.6 and 0.8 ± 0.4, respectively). This effect is due to the reduced dimension of the diameter of the projected spectra distribution on the phasor plot (**Figure 2, Section 1.2**) However, for SNR >= 250, the three methods have similar accuracies, even if the computational time needed for the denoising pre-processing is avoided by applying SuperμMAPPS.

The results regarding a second synthetic P-SHG image of 50x50 pixels, representing a more complex non-uniform pattern resembling experimental data, are reported in **Figure 44**. In this case, each pixel is characterized by a different P-SHG signal simulated with SNR=100 and with the θ_F and γ values color-coded as in **Figure 43**. In this scenario, contiguous pixels are not always characterized by the same parameter values as before, leading to artifacts if the spatial information is considered by the exploited pipeline and denoising methods.

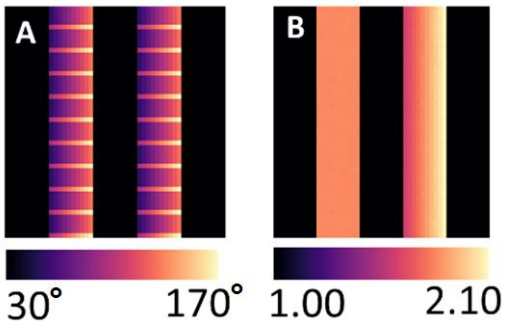

Figure 43: Color-coded maps of θ_F and γ parameters in the non-uniform synthetic image. *Panel **A** shows the θ_F values distribution in the range [30° 170°], while Panel **B** reports the γ distribution with values in the interval [1.00 2.10]. For black pixels $\theta_F = 0°$ and $\gamma = 0$.*

In particular, the vertical bands in **Figure 43A** are characterized by θ_F values varying between 30° and 170° with γ = 1.60, while in **Figure 43B** γ varies in the range [1.00 2.10] with the θ_F distribution shown in **Figure 43A**. Since both the investigated denoising methods act during the pre-processing phase (NOISE2VOID modifies the P-SHG signal, while the median filter is applied to the G_{map} and S_{map} matrices) and the aim of this analysis

is the quantification of the artifacts caused by these approaches compared to SuperμMAPPS, the performance evaluation has been quantified through the G_{map} and S_{map} image spaces.

The true G_{map} and S_{map}, associated to the synthetic P-SHG image, are reported in **Figure 44A** and **Figure 44H** respectively. These images have been considered the Ground Truth, as they are obtained by pixelwise computing the DFT of the simulated P-SHG spectra not affected by noise. These images have been compared with the G_{map} and S_{map} obtained by computing the DFT of the noisy synthetic image (SNR = 100) after the application of NOISE2VOID (**Figure 44D** and **Figure 44K**), Moreover, **Figure 44C** and **Figure 44J** report the G and S matrices obtained by DFT transforming the noisy P-SHG functions and after applying the median-based method (1 iteration). Concerning SuperμMAPPS performance evaluation, the corresponding G_{map} and S_{map} (**Figure 44B** and **Figure 44I**, respectively) have been obtained by pixelwise estimating θ_F and γ from the noisy image and by computing the DFT of the P-SHG spectra simulated (**Eq. 24**) by exploiting the extracted parameters.

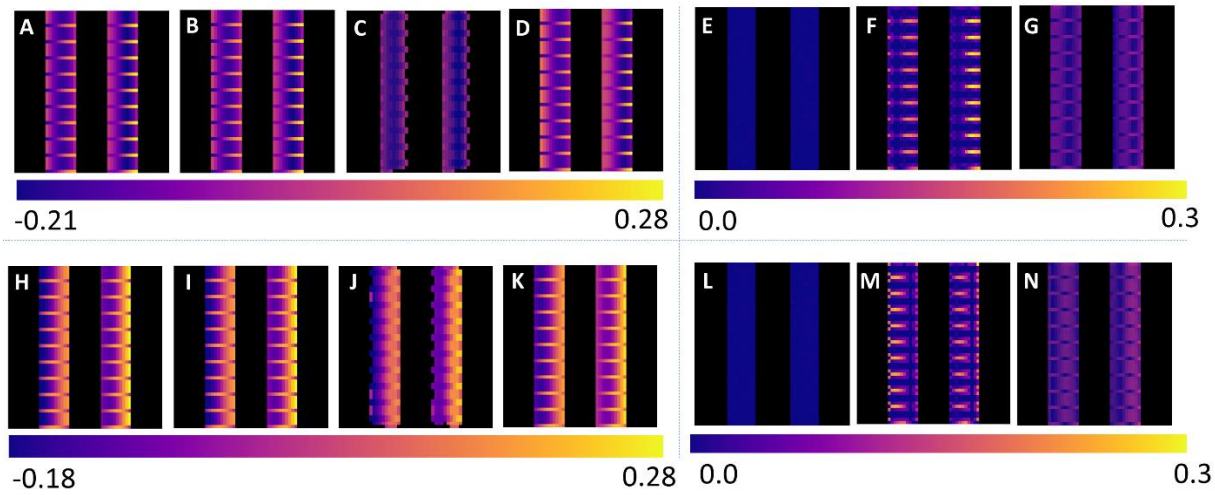

Figure 44: Analysis of the non-uniform synthetic image in terms of G_{map} and S_{map}. *Panels **A** and **H** show the G_{map} and S_{map} obtained by pixelwise computing the DFT of the simulated P-SHG signal The G_{map} and S_{map} related to the PSHG spectra characterized by the θ_F and γ values extracted through SuperμMAPPS, NOISE2VOID and median filter-based pipeline are reported in Panels **B** and **I**, Panels **C** and **J** and Panels **D** and **K**, respectively. The absolute error, computed pixelwise, between the true simulated G_{map} and S_{map} (Panels A*

*and H) and the retrieved values are reported in Panels **E** and **L** (SuperμMAPPS), Panels **F** and **M** (NOISE2VOID) and in Panels **G** and **N** (median filter).*

The absolute error has been computed pixelwise between the G_{map} and S_{map} retrieved by each procedure and the corresponding true simulated values. These results have been provided in **Figure 44E** and **Figure 44L** for the G_{map} and S_{map} associated to SuperμMAPPS, in **Figure 44F** and **Figure 44M** for the median filter-based denoising method and in **Figure 44G** a **Figure 44N** for the pipeline involving NOISE2VOID. In this scenario, SuperμMAPPS outperforms the other approaches in terms of accuracy: indeed, the application of denoising procedures introduced many artifacts, which affect the retrieving of θ_F and γ since they are directly derived from the (G, S) coordinates. In particular, the absolute error associated to each pixel is < 0.01 for both G_{map} and S_{map}, while NOISE2VOID is characterized by discrepancies > 0.04. The worst performance is obtained after the median filter-based denoising, whose differences between the retrieved and the expected (g,s) values are > 0.10 in the boundaries regions, where close pixels are characterized by different simulated P-SHG functions . This effect is due to the fact that both NOISE2VOID and the median filter act by averaging parameters in close pixels, therefore altering their values. Furthermore, this analysis also demonstrates that SuperμMAPPS intrinsically applies a denoising procedure acting directly on the phasor plot by considering an uncertainty circle around the projected point whose radius is inversely proportional to noise. As a consequence, even if SuperμMAPPS has a lower denoising effect than the other investigated procedures (as shown in **Figure 42**), the spatial relation among pixels is not involved, avoiding artifacts in the retrieved G_{map} and S_{map} images.

7.3 Experimental Data Analysis

7.3.1 Experimental Images Acquisition

Here, SuperμMAPPS has been applied in simple experimental conditions to analyze the P-SHG images of murine tendons, whose collagen architecture is simple and well-defined.

These images have been acquired by means of the same two-photon microscopy-based optical setup exploited by Radaelli and colleagues (Radaelli et al. 2017) and the collagen P-SHG signal has been acquired by rotating the incident laser polarization from 0° to 180° in steps of 10°. In this case, in order to reduce the noise affecting the experimental measurement, the images have been downsized from 512x512 to 256x256 pixels. Moreover, only pixels characterized by a signal threshold higher than 150 for each point of the P-SHG curve have been considered and analyzed. Then, images related to human tissue samples (breast carcinoma, skin tissue, thyroid), acquired by a different laboratory and stored in the PSHG-TISS public repository (Hristu et al. 2022), have been processed. In this case, the P-SHG signal coming from different human tissue samples (breast carcinoma, skin tissue, thyroid) have been acquired in both forward and backward directions by exploiting a three-channel Leica TCS SP laser scanning confocal microscope adapted for nonlinear imaging. The incident laser polarization angle θ_L has been changed from 0° to 180° with a step of 20°. Detailed information concerning the acquisition setup have been reported by Hristu and colleagues (Hristu et al. 2022).

7.3.2 Murine tail tendon samples analysis

A preliminary analysis of the SuperμMAPPS performances on experimental data has been conducted on P-SHG images of murine tail tendon samples. Here, SuperμMAPPS results have been compared with the standard pixelwise fitting procedure, which is widely exploited in literature (Tsafas et al. 2020, Paun et al. 2019, Xydias et al. 2021). The results are reported in **Figure 45**.

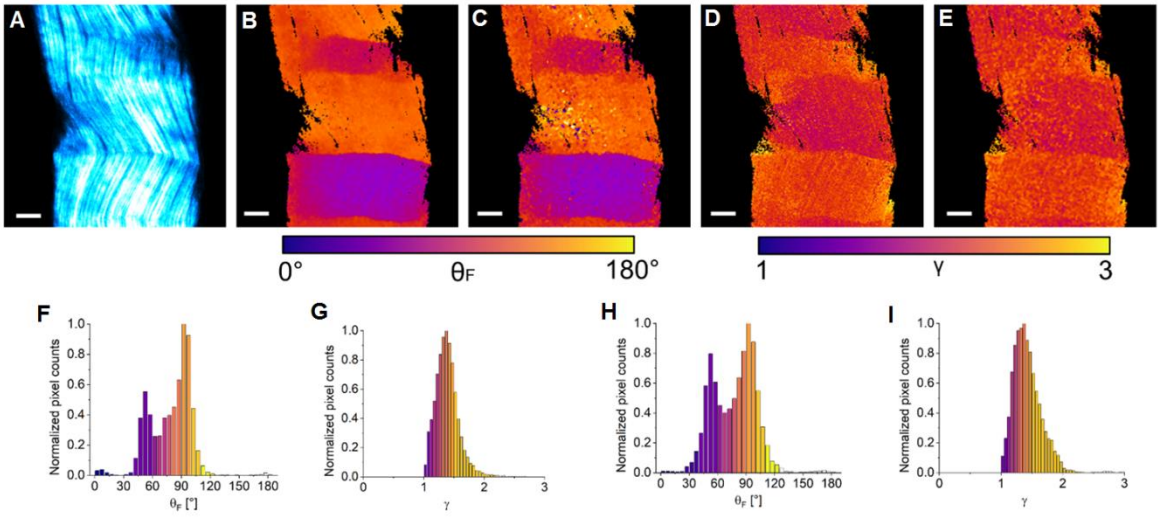

Figure 45: Analysis of a murine tail tendon image. *Panel **A** shows the maximum intensity projection of the acquired P-SHG image stack. Image size: 225x225µm^2 (256x256 pixels). **B and D** are the colormaps referred to the θ_F and γ values pixelwise retrieved by SuperµMAPPS. The color codes for the θ_F and γ values, whose normalized distributions are reported in panels **F** and **G,** respectively. **Panels C and E** represent the colormaps referred to the θ_F and γ values pixelwise retrieved by the standard fitting procedure, while the related normalized distributions are shown in panels **H** and **I.** Scale bars: 30µm.*

Both the colormaps (**Figure 45B-E**) and the histograms (**Figure 45F-I**) highlight that SuperµMAPPS and the standard fitting method (**Eq. 24**) retrieve similar distributions for θ_F and γ. In particular, the θ_F values show two main peaks around 45°-50° and 95°-100°, highlighting different fibrils alignments (**Figure 45A**, **Figure 45B** and **Figure 45C**). Concerning γ, the similarity of the two distributions (**Figure 45G** and **Figure 45I**) is also demonstrated by their fit with a Gaussian function peaked at 1.36 ± 0.17 (SuperµMAPPS) and at 1.39 ± 0.22 (standard fitting method). Both results are compatible with the expected value γ=1.3 for type I collagen.

Then, a second analysis has been performed on the same murine tail tendon sample in order to investigate the effect of NOISE2VOID on experimental data. Due to its poor performances on the synthetic images, the median-based filter technique has not been exploited in this context. NOISE2VOID produces a smoothing effect that can be observed by a comparison between **Figure 45A** and **Figure 46A**, reporting the maximum intensity projection of the analyzed image before and after the application of the denoising method.

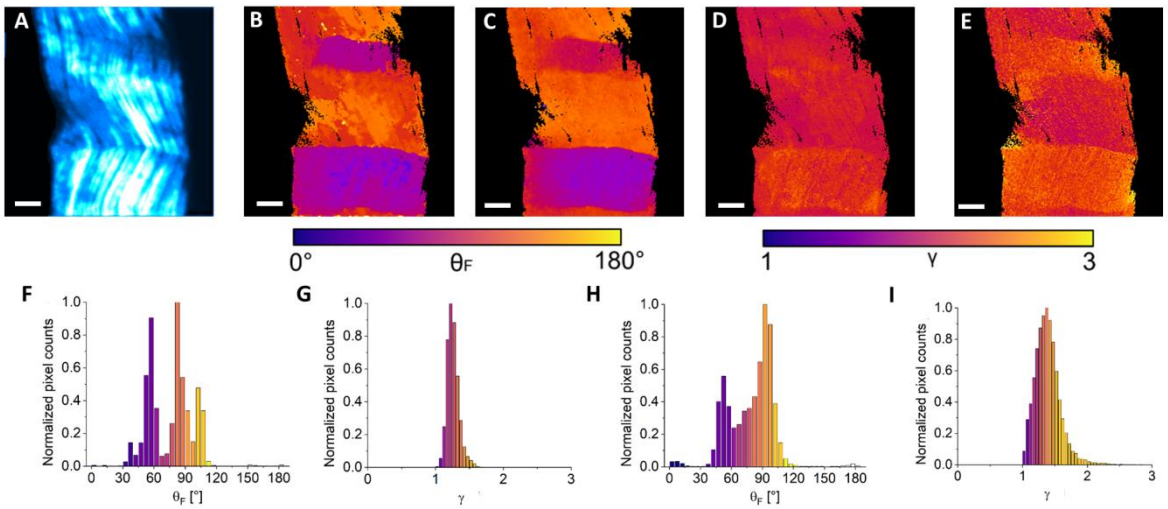

Figure 46: Analysis of a murine tail tendon P-SHG image stack after applying NOISE2VOID. *Panel **A** shows the maximum intensity projection of the acquired P-SHG image stack . Image size: 225x225µm^2 (256x256 pixels). **B, D** are the colormaps referred to the θ_F and γ values pixelwise retrieved by SuperµMAPPS after NOISE2VOID. The color codes for the θ_F and γ values, whose normalized distributions are reported in panels **F** and **G,** respectively. **Panels C and E** represent the colormaps referred to the θ_F and γ values pixelwise retrieved by the standard fitting procedure after NOISE2VOID, while the related normalized distributions are shown in panels **H** and **I**.*

From a qualitative perspective, by observing **Figures 46 B-E**, SuperµMAPPS appears to be more affected by the denoising pre-processing if compared to the standard fitting, as quantitatively demonstrated also by the θ_F and γ values distribution histograms (**Figures 46 F-I**). Indeed, concerning the fitting approach, the two peaks around 45°-50° and 95°-100° characterizing the θ_F distribution are preserved and the γ distribution is peaked at 1.46 ± 0.23, which is not significantly different if compared to the results without NOISE2VOID. Regarding SuperµMAPPS, the γ distribution is sharper and peaked around 1.24 ± 0.08, compatible with the expected results (**Figure 46G**). However, the θ_F histogram is characterized by two main peaks shifted at 50°-55° and 85°-90° with also a third contribution around 105° (**Figure 46F**). This is probably due to the conflict between the denoising procedure of NOISE2VOID and the retrieval phase of SuperµMAPPS. Indeed, NOISE2VOID has been developed to mitigate the effect of white Gaussian noise and it is not effective on structured signals since it modifies the shape of the spectrum (Krull et al. 2020). Unfortunately, SuperµMAPPS is very sensitive to slight modifications of the analyzed spectrum shape, leading to discrepancies in the parameters retrieving.

7.3.3 Comparison with fitting results on human tissue samples

Since the purpose of SuperμMAPPS is providing an effective alternative to the standard PSHG signal analysis, based on fitting procedures, a proper comparison between the results obtained by these two methods has to be provided. In order to accomplish this task, the public repository PSHG-TISS (Hristu et al. 2022) has been exploited. In PSHG-TISS, 240 P-SHG image stacks of 512x512 pixels are collected from different human tissue samples, as reported in **Table 6**. A total of 30 images have been analyzed by means of SuperμMAPPS. In particular, 5 different samples have been processed for 6 clinical cases: 1) H&E-stained normal breast TDLU, 2) H&E-stained in situ breast carcinoma, 3) H&E-stained skin reticular dermis, 4) unstained skin reticular dermis, 5) thyroid papillar carcinoma and 6) thyroid follicular adenoma.

Tissue Type	Staining	Classification	# of patches
Breast	unstained	Normal TDLU	18
		Ductal carcinoma in situ	12
		Invasive carcinoma	17
Breast	H&E	Normal TDLU	18
		Ductal carcinoma in situ	12
		Invasive carcinoma	17
Skin	unstained	papillary dermis	10
		reticular dermis	22
		subcutaneous tissue	21
Skin	H&E	papillary dermis	10
		reticular dermis	22
		subcutaneous tissue	21
Thyroid	H&E	papillary thyroid carcinoma	20
		follicular adenoma	20

Table 6: Tissue types and classification of the 240 images in the PSHG-TISS repository. *The online archive contains three different tissue categories: breast, skin and thyroid. For the breast tissue, samples describing three different growth stages of carcinoma are provided, while the thyroid images are related to a malignant (carcinoma) and a benignant tumor (adenoma). Concerning the skin, regions of papillary dermis, reticular dermis and subcutaneous tissue are collected.*

For each case, the PSHG-TISS repository provides different maps of parameters characterizing the P-SHG signal, obtained by applying the fitting procedure based on **Eq.**

25. Since SuperμMAPPS estimates the θ_F and γ values described in **Eq. 24**, the γ maps related to the fitting procedure (**Figure 48 G-L**) have been obtained by pixelwise computing the ratio χ_{33}/χ_{31}, while θ_F maps (**Figure 47 G-L**) are already provided by the repository. These values have been compared to SuperμMAPPS results in **Figure 47 A-F** and **Figure 48 A-L** for θ_F and γ, respectively. In this scenario, SuperμMAPPS have been applied by exploiting 6 experimental points and 500 permutations since it guarantees reliable retrievals, as previously demonstrated (**Figure 41**). **Table 7** reports the mean difference between the estimated quantities, in terms of absolute value, obtained for θ_F and γ in each considered case.

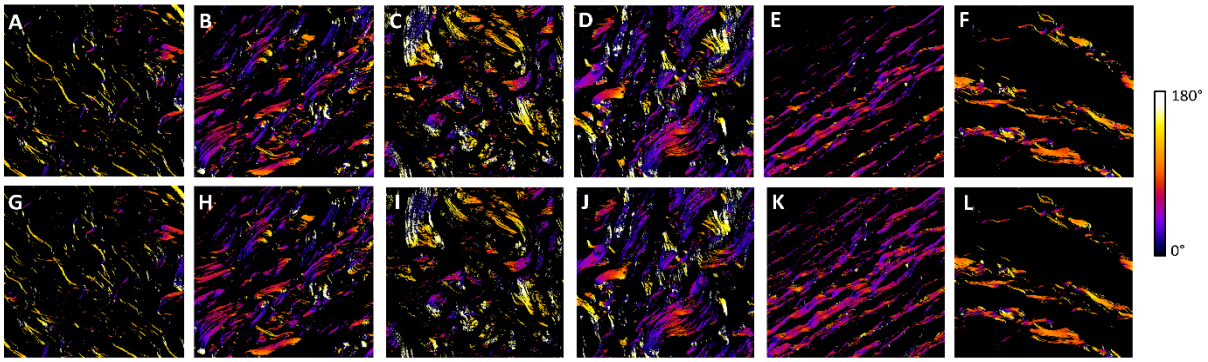

Figure 47: Exemplary θ_F maps obtained by exploiting SuperμMAPPS (A-F) or the fitting procedure (G-L) for different tissues in the PSHG-TISS repository. *The couples **A**, **G** and **B**, **H** show the results obtained by analyzing an H&E-stained normal breast TDLU and an H&E-stained in situ breast carcinoma sample, respectively. Panels **C**, **I** and **D**, **J** are referred to H&E-stained and unstained skin reticular dermis, while panels **E**, **K** and **F**, **L** report the θ_F values obtained by analyzing a thyroid papillar carcinoma and a thyroid follicular adenoma sample, respectively.*

In particular, for each analyzed tissue, the mean discrepancy between the γ values obtained by SuperμMAPPS and the fitting approach is in the range [0.2 0.3]. Regarding the θ_F parameter, higher differences in the fibrils orientation angles provided by SuperμMAPPS and the fitting method are retrieved in the skin tissue (~ 6°-7°), while only ~ 3°-4° of discrepancy are obtained for thyroid and breast tissues. Moreover, the analysis of stained

and unstained skin tissues shows no significant differences, as already demonstrated in the literature (Hristu et al. 2021).

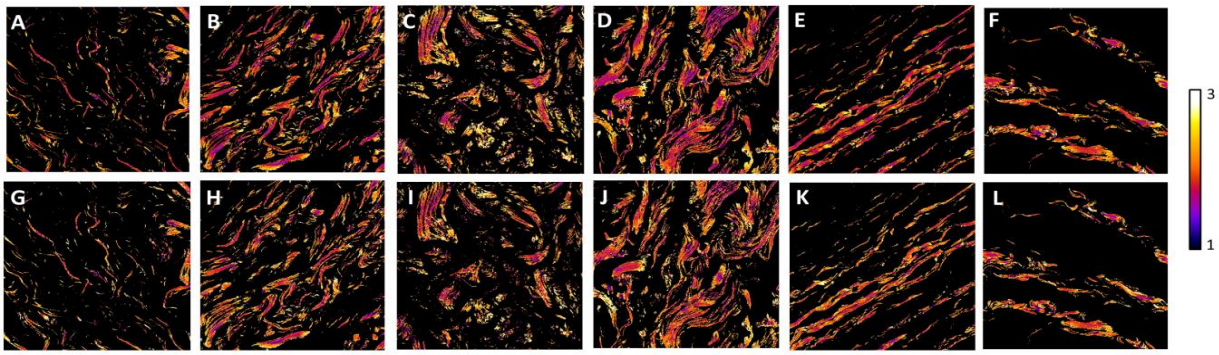

Figure 48: Exemplary γ maps obtained by exploiting SuperμMAPPS (A-F) or the fitting procedure (G-L) for different tissues in the PSHG-TISS repository. *The couples* ***A***, ***G*** *and* ***B***, ***H*** *show the results obtained by analyzing an H&E-stained normal breast TDLU and an H&E-stained in situ breast carcinoma sample, respectively. Panels* ***C***, ***I*** *and* ***D***, ***J*** *are referred to H&E-stained and unstained skin reticular dermis, while panels* ***E***, ***K*** *and* ***F***, ***L*** *report the γ values obtained by analyzing a thyroid papillar carcinoma and a thyroid follicular adenoma sample, respectively.*

This analysis demonstrates that SuperμMAPPS could effectively replace the fitting procedures, which are user-dependent and characterized by higher computational costs, to analyze experimental data in clinical contexts.

Type of tissue	Mean θ_F discrepancy [°]	Mean γ discrepancy
H&E breast normal TDLU	4.1 ± 1.0	0.3 ± 0.2
H&E breast in situ carcinoma	3.7 ± 1.0	0.2 ± 0.2
Unstained skin reticular dermis	7.6 ± 2.5	0.3 ± 0.2
H&E skin reticular dermis	6.4 ± 2.0	0.3 ± 0.2
H&E thyroid papillar carcinoma	2.5 ± 0.9	0.2 ± 0.2
H&E thyroid follicular adenoma	2.4 ± 1.0	0.2 ± 0.2

Table 7: Mean discrepancies between the θ_F and γ values provided by SuperμMAPPS and the fitting approaches, computed for different tissues. *Each result has been obtained by computing the mean value of the parameter differences extracted pixelwise from 5 images for each different clinical case.*

7.3.4 Detection of early stage breast cancer from human samples

Finally, SuperμMAPPS has been tested in a clinical context to provide data useful to assist pathologists in obtaining an early tumor diagnosis. In order to reach this purpose, 47 patches of human breast biopsies (512x512 pixels) have been extracted from the PSHG-TISS public repository (Hristu et al. 2022). In particular, 18 images are classified as normal TDLU (healthy), while 12 and 17 patches refer to early stage in situ carcinoma and invasive carcinoma, respectively. In each image, a density-based clustering technique (Rodriguez et al. 2014) has been applied to the θ_F and γ results retrieved by SuperμMAPPS and the Fibrils Entropy (FE) parameter (Scodellaro et al. 2019) has been computed on the obtained clusters as follows:

$$FE = \frac{-\sum_{i=1}^{N} p_i \log(p_i)}{-\log(\frac{1}{EC})} = \frac{-\sum_{i=1}^{N} \frac{x_i}{EC} \log(\frac{x_i}{EC})}{-\log(\frac{1}{EC})} \quad (37)$$

where p_i is the probability of occurrence of the i-th cluster measured as the ratio between the number of elements in the i-th cluster and the total number of clustered elements EC, and N is the total number of clusters. For a perfectly ordered microstructure, a single cluster, N = 1, and $p_i = 1$ are expected, therefore FE = 0. For a maximally disorder microstructure, instead, N = EC, and $p_i = 1/N$ are expected. In this case, FE = 1.

The distributions of the FE values and of the total number of clusters for healthy, early stage and invasive carcinoma stages have been reported in **Figure 49A** and in **Figure 49B,** respectively. The clustering method has been applied by considering a density threshold of 10% for θ_F and γ (it means that cluster centroids are expected to have at least the 10% of the data in their neighborhood in both θ_F and γ feature spaces), and by removing the clusters

characterized by few elements (< 2% of the analyzed image pixels). A two samples t-test analysis has been performed to highlight differences between the healthy and tumorous distributions.

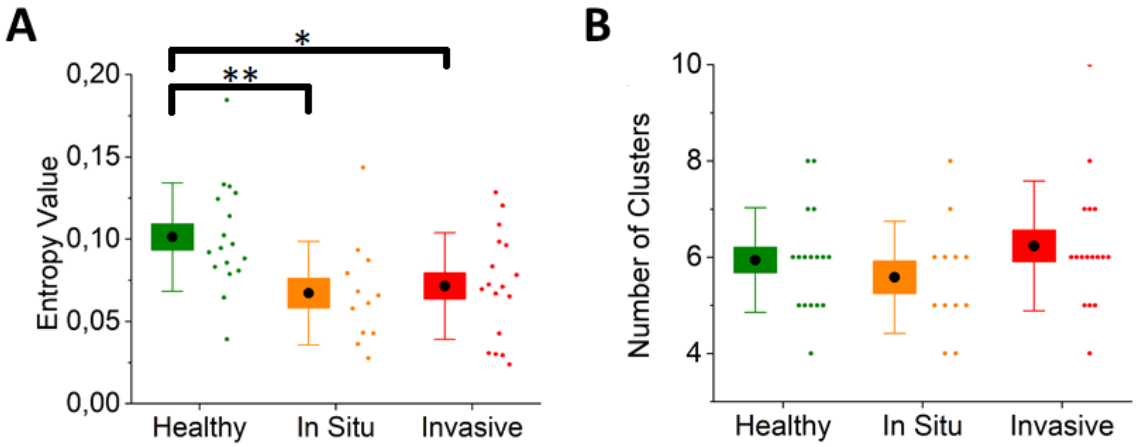

Figure 49: t-test results of the SuperμMAPPS analysis performed on human breast cancer samples. *The distributions of the Fibrils Entropy and the number of clusters for healthy (green), early stage (orange) and invasive (red) human breast carcinomas are reported in panels **A** and **B**, respectively, both in scatterplot (right) and in boxplot (left) format. A density threshold of 10% for θ_F and γ has been considered during the clustering procedure. The black dot represents the mean value of each distribution, while the bars highlight the standard deviation. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

As highlighted in **Figure 49B**, no significant differences have been detected for the number of clusters distribution. Conversely, the t-test analysis underlined a significant discrepancy between the FE values related to the healthy and breast cancer regions. In particular, a p value < 0.01 has been detected between the healthy and the early stage distributions, while a p value < 0.05 has been computed between the healthy and the invasive tumor biopsies FE values. These results demonstrate the capability of SuperμMAPPS to successfully identify early stage tumor development. Moreover, the results, which show a more ordered collagen microstructure (i.e. a lower entropy value) for tumor biopsies, are compatible with previous studies (Wu et al. 2018). Other clustering configurations have been analyzed while varying the density threshold from 2% to 20% and by fixing the dimension of the smallest cluster to 2% (as recommended by the developers, Rodriguez et al. 2014), as reported in **Figure 50**.

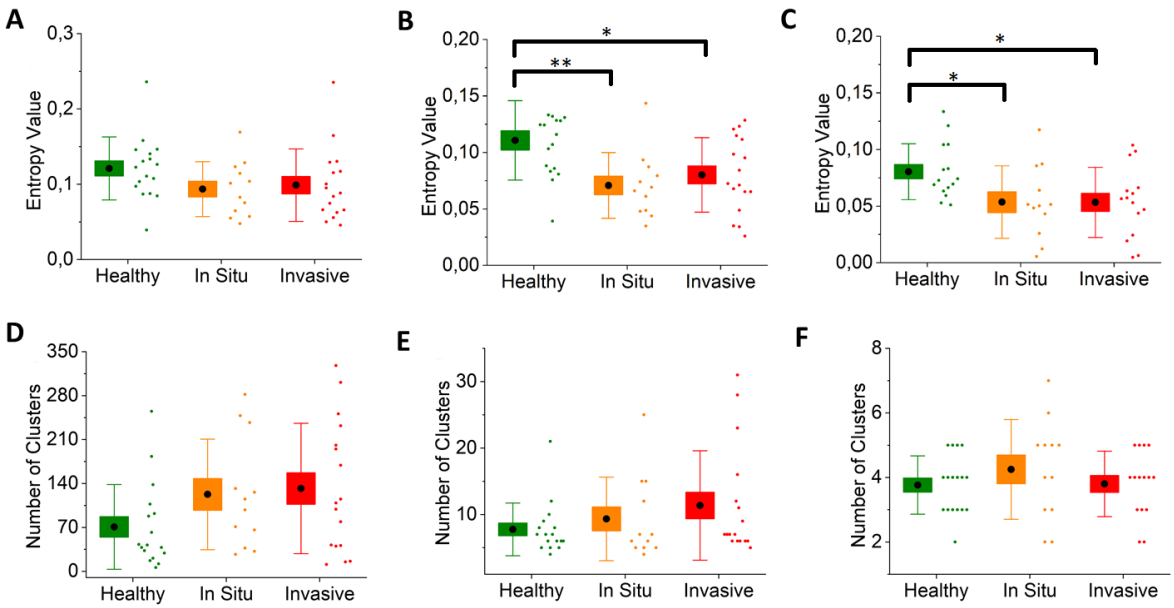

Figure 50: t-test results of the SuperμMAPPS analysis performed on human breast cancer samples by varying the clustering configurations. *The distributions of the Fibrils Entropy and the number of clusters for healthy (green), early stage (orange) and invasive (red) human breast carcinomas are reported both in scatterplot (right) and in boxplot (left) format for each configuration. In panels* ***A*** *and* ***D****, a density threshold of 2% for* θ_F *and* γ *have been considered during the clustering procedure. Different thresholds have been selected for panels* ***B*** *and* ***E*** *(5%) and panels* ***C*** *and* ***F*** *(20%). The black dot represents the mean value of each distribution, while the bars are associated to their standard deviation. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

Opposite to the number of clusters, FE values show significant discrepancies between healthy and tumor regions for density thresholds between 5% and 20% (**Figure 50B** and **Figure 50C**). This characterization demonstrates the robustness of the proposed approach since the significance of the FE value in detecting early tumor stage is not affected by minimal changes of the clustering hyperparameters. Moreover, the number of clusters is not suitable to perform tissue classification for any clustering configuration (**Figures 50E** and **Figures 50F**).

8. Discussion and Future Perspectives

During my PhD project, I developed three different algorithms based on the coupling of the standard phasor approach with different machine learning techniques. For the first time, the phasor approach has been successfully coupled to supervised machine learning-based protocols to perform regression tasks, paving the way for novel applications of this method. Differently from the most recent applications of the phasor approach (Vallmitjana et al. 2021), which implement simple clustering procedure in the phasor space, the proposed pipelines are trained to associate to each coordinate (g,s) of the phasor plot a set of values which reflect one of the possible combination of the involved parameters describing the shape of the analyzed spectrum. This goal has been achieved by exploiting a huge training set based on numerical models, able to include a wide range of parameters configurations affecting the shape of the analyzed signal. Indeed, each proposed algorithm is a general pipeline of spectra/signal analysis that can be easily extended to other applications, under the condition that the aim of the analysis is the retrieving of parameters whose values modify the shape of the functions analyzed by the phasor approach. Moreover, another crucial novelty of the three proposed approaches concerns the handling and evaluation of noise affecting the experimental measurements, as highlighted in the following.

Concerning i-φ-MaLe, this algorithm has been developed to simultaneously and consistently retrieve different biophysical parameters, full F spectra, fluorescence corrected for reabsorption and fluorescence quantum yield by exploiting the DFT of the R_{app} spectrum. The validation and the retrieving accuracy quantification of i-φ-MaLe has been performed by exploiting a detailed analysis, involving every hyperparameter characterizing the method (spectral windows width, number of exploited phasor planes, size of the dataset), white noise usually affecting experimental data, and the estimation of seasonal and daily trends of each biophysical and F parameter on two widely exploited benchmarks: crops from the top of canopy level and a deciduous forest from tower observations. i-φ-MaLe represents a crucial novelty in the remote sensing field, as biophysical parameters and full F spectra are

retrieved simultaneously for the first time. The robustness of the algorithm is assured by the set of simulations exploited to build the training dataset, which is based on the most recent and well-known atmosphere-canopy radiative transfer model MODTRAN+SCOPE. However, also the accuracy between the retrieved values and the experimental measurements depends on the effective representativity of MODTRAN+SCOPE pipeline for the investigated canopy. As a consequence, other canopy-related quantities, as the Leaf Inclination Distribution Function (LIDF), have to be properly set to improve the capability of i-φ-MaLe to analyze different types of vegetation. However, the same consideration can be extended to the state-of-the-art model inversion methods (Celesti et al. 2018, Verhoef et al. 2019), devoted to the analysis of R_{app} spectra. Moreover, the i-φ-MaLe accuracy tested on radiative transfer simulations is comparable with other fluorescence retrieval methods: it is slightly lower for F_{760} (RRMSE=1% vs. RRMSE=0.5% for SpecFit) and larger for F_{687} (RRMSE=1% vs. RRMSE=2% for SpecFit). Concerning the algorithm computational cost, the overall time is about 3 ms per spectrum even for the largest size of the training data set used here (200000 points) on a 16 core CPU, 128 GB RAM machine. Such a reduced value of CPU time would allow to implement the parameter retrieval online during the R_{app} acquisitions that are typically sampled at several minutes cadency. This paves the way to the possibility to increase the complexity of i-φ-MaLe, for instance by exploiting the more accurate 3D radiative transfer models for the training dataset implementation. ML approaches to the analysis of spectral data from vegetation have been reported in the literature (Verrelst et al., 2019) and they are mainly devoted to the retrieval of biophysical parameters. However, they usually limit their analysis to one or two parameters, in particular LAI and Cab (De Grave et al. 2020, Bhadra et al. 2020), while i-φ-MaLe is able to estimate not only a wider set of biophysical parameters, including the APAR and the Fqe, but also the fluorescence spectra, nowadays retrieved by exploiting inversion-based algorithms (Cogliati et al. 2019). Moreover, i-φ-MaLe has been able to estimate, from real data, the fluorescence quantum yield Fqe and the fluorescence spectrum at photosystem level F_{RC}, parameters which are experimentally retrieved for the first time in literature. At the state of

the art, none studies concerning the development of Machine Learning algorithm for apparent reflectance data analysis has been able to partition the R_{app} spectrum and provide a quantitative description of the most relevant spectral windows for each variable (**Figure 15**), even if this is crucial to select the most informative wavelengths to analyze in scenarios where the absorption effect of the oxygen bands is relevant, as tower-level or satellite-level measurements. In particular, as expected, the red-edge region is the most important to retrieve Cab, red and near-infrared to quantify LAI and also APAR is relevant in the O_2-A band. F and F_{RC} are better detected inside the O_2 bands, while interestingly the O_2-B resulted as the most important spectral region to retrieve Fqe, confirming that red F offers most relevant information about F emission efficiency. The good compatibility between the seasonal and daily trends for top of canopy-level experimental data retrieved by i-*φ*-MaLe, SFM and SpecFit (**Figure 21**) further demonstrates the capability of the proposed approach to provide robust results. Moreover, the retrieval performance analysis obtained by excluding the O_2 bands and the compatibility between the LAI estimated by i-*φ*-MaLe and the LAI measured on the field extend the applicability of the method to tower and satellite-level measurements (**Figure 27** and **Figure 28**). Indeed, this analysis clearly shows that i-*φ*-MaLe is able to perform F estimations also in scenarios where the telluric oxygen atmospheric absorption affects the R_{app} measurements, differently from SFM and SpecFit, which are totally dependent on the information content coming from the O_2 bands. Finally, the preliminary GPP retrieval offered by i-*φ*-MaLe (**Figure 26**) leads to further extension of the method to extract novel parameters.

A demonstration of the wide range of applications of i-*φ*-MaLe is represented by SuperμMAPPS, developed to analyze the P-SHG signal of collagen in order to quantitatively describe its microarchitecture. In particular, SuperμMAPPS processes image stack, where each pixel is associated to a P-SHG signal. SuperμMAPPS builds on the i-Φ-male algorithm by exploiting permutations of the P-SHG signal to apply the statistical-based retrieval algorithm. In particular, good performances (MAE < 2° for θ_F and < 0.2 for γ) have

been obtained by evaluating 500 phasor plots (**Figure 41**). This means that only 6 experimental points are sufficient to apply this algorithm, implying a 65% acquisition time reduction of each P-SHG image stack. Similarly to i-φ-MaLe, SuperμMAPPS algorithm has been evaluated in dependence of its hyperparameters (the number of exploited permutation, i.e. the number of phasor plots) and computational time (~7ms per spectrum), resulting a really accurate algorithm, characterized by low uncertainties (0.6° ± 2.0° for the θ_F parameter and 0.10 ± 0.09 for γ when more than 250 different permutations and SNR = 100 are considered). Differently from other already published denoising methods (Cutrale et al. 2017, Krull et al. 2020), which consider the spatial encoding of the P-SHG signal (both in the image and in the G, S matrices in the Fourier space), SuperμMAPPS evaluates the noise effect directly in the phasor space without taking into account the spatial relation or any mathematic operation among close pixels. This is a great improvement in phasor-based analysis, which paves the way towards future studies concerning the development of novel and more accurate pipelines devoted to the denoising and reconstruction of signals. The versatility of SuperμMAPPS has been demonstrated by its application to different tissues, both murine (tail tendon, **Figure 45** and **Figure 46**), and human (breast, skin, thyroid, **Figure 47** and **Figure 48**). In particular, the analysed human samples have been extracted from the PSHG-TISS public repository: the θ_F and γ values provided by SuperμMAPPS have been compared to the standard fit results obtained by the PSHG-TISS authors. The strong compatibility with the quantifications performed by an independent research group and the possibility to have free access to the analysed images is a really important achievement to guarantee the data reproducibility and describe SuperμMAPPS as an algorithm conceived by following the Open Science approach. Moreover, the fibrils entropy parameter (Scodellaro et al. 2019), computed through the clustering analysis of θ_F and γ extracted by SuperμMAPPS, has been demonstrated to be a useful novel pipeline to automatically recognize breast cancer also in its early stage, hardly detected in histopathology by exploiting standard imaging and analysis techniques as the semi-quantitative evaluation of fixed stained samples under the microscope. Indeed, if compared

to the healthy tissue, SuperμMAPPS highlighted that both early stage and full growth breast cancer samples are characterized by a more ordered collagen micro-structure (i.e. a lower value of the fibril entropy parameter, **Figure 49**), also confirming recent published studies (Wu et al. 2018).

Nowadays, one of the most crucial issues in applying artificial intelligence to histopathological analysis is due to the color discrepancy of the same biological structure in tissue sections stained and acquired under different experimental conditions or in different laboratories. In this scenario, Φ-Norm has been demonstrated a valid pre-processing tool, able to improve the accuracy of machine learning techniques which analyze the image color content. This proposed pipeline is the first method able to perform the stain normalization by exploiting the geometrical properties of the phasor approach. Here, we reported the comparison of Φ-Norm with other published methods (Macenko et al., 2009, Reinhard et al. 2001) regarding the stain normalization performance. The visual inspection of the images highlighted the compatibility among the three staining normalization methods. It demonstrated the good accuracy of Φ-Norm, able to obtain results comparable to two of the most exploited protocols, which are considered the state of the art. This consideration has been also supported by the quantitative characterization of the algorithm performance in terms of the red, green and blue channels intensity distributions and the 3D visualization in the RGB colorspace. This analysis highlighted that the histograms related to the Macenko, Reinhard and Φ-Norm methods are superimposed to the reference image distribution. Moreover, Reinhard and Φ-Norm better match the reference image distributions for the green and blue channel, respectively. The 3D histogram visualization easily highlights slight discrepancies between the color transformation performed by Φ-Norm, Macenko and Reinhard approaches. For instance, the color expression of the collagen in the PicroSirius Red staining, which is different between Reinhard and the other approaches, is represented by a different 3D distribution.

Furthermore, despite its very simple structure, Φ-Norm has been demonstrated to be a general method, able to perform color normalization also by considering other staining, as the PicroSirius Red (**Figure 34, Figure 38**).

Finally, the color content of biological structures manually segmented in stain normalized images has been quantitatively characterized by comparing their 3D distributions and by computing a novel metric, named Color Mean Distance (CMD), which is the difference between the mean color intensity of each considered structure among the 20 biopsies and the mean color intensity of the structure in the reference image. The lower CMD value obtained by means of Φ-Norm with respect to the other stain normalization algorithms quantitatively demonstrates that our method is the most suitable in order to obtain a uniform color describing the same biological structure. In particular, compared to Macenko, the discrepancies in terms of color content from the reference values are reduced by ~50% for nuclei and ~40% for cells cytoplasm in H&E patches. Concerning the SR images, the Φ-Norm color discrepancy is reduced by ~40% for nuclei and by ~25% for collagen, compared to the Reinhard procedure. This leads to the possibility to exploit Φ-Norm as a future standard approach in the pre-processing pipelines of artificial intelligence algorithms designed for histopathology. Indeed, Φ-Norm has been also applied to entire SR-stained whole slide images (**Figure 38**). In this case, the application of color-based algorithms is really challenging due to the color content discrepancy in different part of the images (in our case, dark orange in the centre of the sample and yellow near the boundaries). Φ-Norm performed an accurate staining normalization of the entire section, as demonstrated by ~55% reduction in the CMD computed for the cells cytoplasm with respect to the original acquired image. The same procedure has been also applied to analyze entire H&E-stained whole slide sample (**Figure 39**), obtaining a ~80% reduction in the CMD computed for the cells cytoplasm with respect to the original acquired image. These results imply an important achievement for the histopathology field since Φ-Norm is able to strongly reduce color expression discrepancies caused by different staining procedures, which highly affect

the reliability of the expert diagnosis, and can be easily exploited by pathologists with standard computational capabilities.

In summary, these studies demonstrate the possibility to boost the geometrical and mathematical properties of the phasor approach by exploiting supervised machine learning techniques. The application of the three proposed phasor-based tools, i-φ-MaLe, SuperμMAPPS and Φ-Norm, in different research fields proves the outstanding versatility of the novel pipeline I conceived during my PhD project. In particular, these studies pave the way to new research fields concerning the phasor-based signal and image analysis by breathing new life into the standard phasor approach.

References

Adnan MK. et al. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Transactions on Biomedical Engineering. https://doi.org/ 10.1109/TBME.2014.2303294

Amat-Roldan I. et al. (2010). Fast image analysis in polarization SHG microscopy. Opt. Express. https://doi.org/10.1364/OE.18.017209.

Anghel, A. et al. (2019). A high-performance system for robust stain normalization of whole-slide images in histopathology. Frontiers in medicine. https://doi.org/10.3389/fmed.2019.00193.

Baker, N. R. (2008). Chlorophyll fluorescence: a probe of photosynthesis in vivo. Annu. Rev. Plant Biol, 59, 89–113. doi: 10.1146/annurev.arplant.59.032607.092759.

Bakheet, S. et al. (2021). Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification. Computers in Biology and Medicine. https://doi.org/10.1016/j.compbiomed.2021.104781.

Bella J. et al. (1994). Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. Science. https://doi.org/10.1126/science.7695699.

Berk, A. et al. (2011). MODTRAN 5.2.1 User's Manual. 4 Fourth Ave., Burlington, MA 01803-3304, Air Force Research Laboratory, Space Vehicles Directorate, Air Force Materiel Command, Hanscom AFB, MA 01731- 3010: Spectral Sciences, Inc.

Bhadra, S. et al. (2020). Quantifying Leaf Chlorophyll Concentration of Sorghum from Hyperspectral Data Using Derivative Calculus and Machine Learning. Remote Sensing. https://doi.org/10.3390/rs12132082.

Bi, W.L. et al. (2019). Artificial Intelligence in cancer imaging : clinical challenges and applications. Cancer journal for clinicians. https://doi.org/10.3322/caac.21552.

Buehler, MJ. (2006). Nature designs tough collagen: explaining the nanostructure of collagen fibrils. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.0603216103

Bug, D. et al. (2015). Foreground extraction for histpathological whole slide imaging. Bildverarbeitung für die Medizin, Springer.

Burova, AY. (2021). Concept of multistage discrete fourier transform without performing multiplications. J. Phys.: Conf. Ser. 1889 022003.

Campagnola PJ. et al. (2002). Three-dimensional high-resolution second-harmonic generation imaging of endogenous structural proteins in biological tissues. Biophysical Journal. https://doi.org/ 10.1016/S0006-3495(02)75414-3.

Campagnola PJ. et al. (2012). Second harmonic generation microscopy: principles and applications to disease diagnosis. Laser and Photonics Reviews. https://doi.org/10.1002/lpor.200910024

Campbell, P. K. E. et al. (2019). Diurnal and Seasonal Variations in Chlorophyll Fluorescence Associated with Photosynthesis at Leaf and Canopy Scales. Remote Sensing, https://doi.org/10.3390/rs11050488.

Cardiff, RD. Et al. (2014). Manual hematoxylin and eosin staining of mouse tissue sections. Cold Spring Harbor Protocols. pdb-prot073411.

Cawse-Nicholson, K. et al. (2021). NASA's surface biology and geology designated observable: A perspective on surface imaging algorithms. Remote Sensing of Environment.

https://doi.org/10.1016/j.rse.2021.112349.

Celesti, M. et al. (2018). Exploring the physiological information of Sun-induced chlorophyll fluorescence through radiative transfer model inversion. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2018.05.013.

Cendrero-Mateo, MP. et al. (2019). Sun-Induced Chlorophyll Fluorescence III: Benchmarking Retrieval Methods and Sensor Characteristics for Proximal Sensing. Remote Sensing. https://doi.org/10.3390/RS11080962.

Chang, CY et al. (2020). Systematic assessment of retrieval methods for canopy far- red solar- induced chlorophyll fluorescence (SIF) using high-frequency automated field spectroscopy. Journal of Geophysical Research: Biogeosciences. https://doi.org/10.1029/2019JG005533.

Chapeau-Blondeau, F. et al. (2020). Fourier-transform quantum phase estimation with quantum phase noise. Signal Processing. https://doi.org/10.1016/j.sigpro.2019.107441.

Chernov V. et al. (2015). Integer-based accurate conversion between RGB and HSV color spaces. https://doi.org/10.1016/j.compeleceng.2015.08.005.

Cicchi, R. et al. (2013). From molecular structure to tissue architecture: collagen organization probed by SHG microscopy. J. Biophotonics. https://doi.org/10.1002/jbio.201200092

Clark, K. et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging. https://doi.org/10.1007/s10278-013-9622-7 .

Cogliati, S. et al. (2015a). Continuous and long-term measurements of reflectance and sun-induced chlorophyll fluorescence by using novel automated field spectroscopy systems. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2015.03.027.

Cogliati, S. et al. (2015b). Retrieval of sun-induced fluorescence using advanced spectral fitting methods. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2015.08.022.

Cogliati, S. et al. (2018). Red and Far-Red Fluorescence Emission Retrieval from Airborne High-Resolution Spectra Collected by the Hyplant-Fluo Sensor. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. https://doi.org/10.1109/IGARSS.2018.8517758.

Cogliati, S. et al. (2019). A Spectral Fitting Algorithm to Retrieve the Fluorescence Spectrum from Canopy Radiance. Rem. Sens., https://doi.org/10.3390/rs11161840.

Coudray, N. et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine. https://doi.org/10.1038/s41591-018-0177-5

Cutrale, F. et al. (2017). Hyperspectral phasor analysis enables multiplexed 5D in vivo imaging. Nat Methods, https://doi.org/10.1038/nmeth.4134.

De Grave, C. et al. (2020). Quantifying vegetation biophysical variables from the Sentinel-3/FLEX tandem mission: Evaluation of the synergy of OLCI and FLORIS data sources. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2020.112101.

Diaspro, A. et al. (2006). Two-photon fluorescence excitation and related techniques in biological microscopy. https://doi.org/ 10.1017/S0033583505004129.

Digman, M.A. et al. (2008). The Phasor Approach to Fluorescence Lifetime Imaging Analysis. Biophys. J., https://doi.org/10.1529/biophysj.107.120154.

Duhamel P. et al. (1990). Fast fourier transforms: A tutorial review and a state of the art. Signal processing. https://doi.org/10.1016/0165-1684(90)90158-U.

Fereidouni, F. et al. (2012). Spectral phasor analysis allows rapid and reliable unmixing of fluorescence microscopy spectral images. Opt Express, https://doi.org/10.1364/OE.20.012729.

Fischer, AH, et al. (2008). Hematoxylin and eosin staining of tissue and cell sections. Cold spring harbor protocols. pdb-prot4986.

Frankenberg, C. et al. (2018). The Chlorophyll Fluorescence Imaging Spectrometer (CFIS), mapping far red fluorescence from aircraft. Remote Sensing of Environment. https://doi.org/10.1016/J.RSE.2018.08.032.

Gastellu-Etchegorry JP. Et al. (2017). DART: Recent Advances in Remote Sensing Data Modeling With Atmosphere, Polarization, and Chlorophyll Fluorescence. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2017.2685528.

Goppert-Mayer M. (1931). Uber elementarakte mit zwei quantensprunngen. Ann. Phys. (Leipzig), 9:1273–295.

Grossmann, K.. et al. (2018). PhotoSpec: A new instrument to measure spatially distributed red and far-red Solar-Induced Chlorophyll Fluorescence. Remote Sensing of Environment. https://doi.org/10.1016/J.RSE.2018.07.002.

Gurcan MN. Et al. (2009). Histopathological image analysis: a review. Biomed Eng IEEE Rev. https://doi.org/10.1109/RBME.2009.2034865.

Hanumantharaju, MC. et al. (2011). A novel FPGA based reconfigurable architecture for image color space conversion. International Conference on Computing and Communication Systems. https://doi.org/ 10.1007/978-3-642-29216-3_32

Haralick, RM. Et al. (1973) "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics. https://doi.org/10.1109/TSMC.1973.4309314

Hristu, R. et al. (2016). Improved quantification of collagen anisotropy with polarization-resolved second harmonic generation microscopy. Journal of Biophotonics. https://doi.org/10.1002/jbio.201600197.

Hristu, R. et al. (2021). Influence of hematoxylin and eosin staining on the quantitative analysis of second harmonic generation imaging of fixed tissue sections. Biomed. Opt. Expr. https://doi.org/10.1364/BOE.428701.

Hristu, R. et al. (2022). PSHG-TISS: A collection of polarization-resolved second harmonic generation microscopy images of fixed tissues. Sci Data. https://doi.org/10.1038/s41597-022-01477-1

Jameson DM. et al. (1984). The Measurement and Analysis of Heterogeneous Emissions by Multifrequency Phase and Modulation Fluorometry. Applied spectroscopy reviews. https://doi.org/10.1080/05704928408081716

Jothi, N. et al. (2015). Data Mining in Healthcare – A Review. Procedia Computer Science. https://doi.org/10.1016/j.procs.2015.12.145.

Kaushik CSH et al. (2014). A tutorial review on discrete fourier transform with data compression application. *International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE),* https://doi.org/10.1109/ICGCCEE.2014.6922210.

Khan A.M. et al. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Transactions on Biomedical Engineering. https://doi.org/ 10.1109/TBME.2014.2303294.

Kleczek, P. et al. (2020). A novel method for tissue segmentation in high-resolution H&E-stained histopathological whole-slide images. Computerized Medical Imaging and Graphics.

https://doi.org/10.1016/j.compmedimag.2019.101686.

Krull A. et al. (2020) Probabilistic Noise2Void: Unsupervised Content-Aware Denoising. Front. Comput. Sci. https://doi.org/10.3389/fcomp.2020.00005.

Kryjak et al. (2014). Real-time background generation and foreground object segmentation for high-definition colour video stream in FPGA device. Journal of Real-Time Image Processing. https://doi.org/ 10.1007/s11554-012-0290-5.

Kutay, MA. et al. (1998). Optimal image restoration with the fractional Fourier Transform. JOSA A. https://doi.org/10.1364/JOSAA.15.000825

Lacomb R. et al. (2008). Quantitative Second Harmonic Generation Imaging of the Diseased State Osteogenesis Imperfecta: Experiment and Simulation. Biophysical Journal. https://doi.org/10.1529/biophysj.107.114405

Lakshmanan, BS. Et al. (2019). Stain removal through color normalization of haematoxylin and eosin images: A review. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1362/1/012108

Lattouf, R. et al. (2014). Picrosirius red staining: a useful tool to appraise collagen networks in normal and pathological tissues. J Histochem Cytochem. https://doi.org/ 10.1369/0022155414545787.

Liao, YH. et al. (2013). Determination of chronological aging parameters in epidermal keratinocytes by in vivo harmonic generation microscopy. Biomed. Opt. Express. https://doi.org/ 10.1364/BOE.4.000077

Lopez, M. G., et al. (2012). BCDR: a breast cancer digital repository. 15th International conference on experimental mechanics.

Lu, X. et al. (2020). Comparison of total emitted solar-induced chlorophyll fluorescence (SIF) and top-of-canopy (TOC) SIF in estimating photosynthesis. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2020.112083.

Luo W. et al. (2016). Propagation phasor approach for holographic image reconstruction. Scientific Reports. https://doi.org/10.1038/srep22738

Macenko M. et al. (2009). A method for normalizing histology slides for quantitative analysis. IEEE International Symposium on Biomedical Imaging. https://doi.org/10.1109/ISBI.2009.5193250.

Malacrida, L. et al. (2021). The phasor plot. A universal circle to advance fluorescence lifetime analysis and interpretation. Ann. Rev. of Biophys. https://doi.org/10.1146/annurev-biophys-062920-063631.

Malenovsky, Z. et al. (2009). Scientific and technical challenges in remote sensing of plant canopy reflectance and fluorescence. Journal of experimental botany. https://doi.org/10.1093/jxb/erp156

Mazzoni, M. et al. (2010). High-resolution methods for fluorescence retrieval from space. Optics Express. https://doi.org/10.1364/OE.18.015649.

Mazumder N. et al. (2019). Label-Free Non-linear Multimodal Optical Microscopy—Basics, Development, and Applications. Frontiers in Physics. https://doi.org/10.3389/fphy.2019.00170

Meroni, M. et al. (2009). "Remote sensing of solar-induced chlorophyll fluorescence: Review of methods and applications. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2009.05.003

Meroni, M. et al. (2010). Performance of Spectral Fitting Methods for vegetation fluorescence quantification. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2009.09.010.

Mitchell, T. (1990). Machine Learning. Annual Review on Computer Science. https://doi.org/10.1146/annurev.cs.04.060190.002221.

Mohammed, G. H. et al. (2019). Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50 years of progress. Rem. Sens. Environ., https://doi.org/10.1016/j.rse.2019.04.030.

Mohanaiah, P. et al. (2013). Image texture feature extraction using GLCM approach. International journal of scientific and research publications. Int J Sci Res Publ 3(5) (ISSN: 2250-3153).

Nemat, H. et al. (2018) "Classification of breast lesions in ultrasonography using sparse logistic regression and morphology-based texture features." Medical physics. https://doi.org/10.1002/mp.13082.

Oheim M. et al. (2006). Principles of two-photon excitation fluorescence microscopy and other nonlinear imaging approaches. Advanced Drug Delivery Reviews. https://doi.org/10.1016/j.addr.2006.07.005

Pantanowitz, L. et al. (2011). Review of the current state of whole slide imaging in pathology. Journal of Pathology Informatics. https://doi.org/10.4103/2153-3539.83746.

Parisi, GI. Et al. (2019). Continual lifelong learning with neural networks: A review. Neural networks. https://doi.org/10.1016/j.neunet.2019.01.012.

Paun, B. et al. (2019). Strategies for Optimizing the Determination of Second-Order Nonlinear Susceptibility Tensor Coefficients for Collagen in Histological Samples. IEEE Access. https://doi.org/ 10.1109/ACCESS.2019.2937360.

Pinto, F. et al. (2020). Dynamics of sun-induced chlorophyll fluorescence and reflectance to detect stress-induced variations in canopy photosynthesis. Plant, Cell & Environ. https://doi.org/10.1111/pce.13754.

Psilodimitrakopoulos S. et al. (2016). Polarization second harmonic generation discriminates between fresh and aged starch-based adhesives used in cultural heritage. Microsc. Microanal. https://doi.org/ 10.1017/S1431927616011570

Rabinovich, A. et al. (2003). Unsupervised color decomposition of histologically stained tissue samples. Adv. in Neural Inf. Proc. Systems.

Radaelli, F. et al. (2017). µMAPPS: a novel phasor approach to second harmonic analysis for in vitro-in vivo investigation of collagen microstructure. https://doi.org/10.1038/s41598-017-17726-y

Raja, AM. et al. (2015). Differential remodeling of extracellular matrices by breast cancer initiating cells. J. Biophotonics. https://doi.org/ 10.1002/jbio.201400079

Rajabi, E. et al. (2022). A structured review of sparse fast Fourier transform algorithms. Digital Signal Processing. https://doi.org/10.1016/j.dsp.2022.103403.

Reddy KA. et al. (2008) Use of Fourier Series Analysis for Motion Artifact Reduction and Data Compression of Photoplethysmographic Signals. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/TIM.2008.2009136.

Reinhard, E. et al. (2001). Color transfer between images. IEEE Computer graphics and applications https://doi.org/10.1109/38.946629.

Rivenson Y. et al. (2019). PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light: Science & Applications. https://doi.org/10.1038/s41377-019-0129-y.

Rivenson Y. et al. (2020). Emerging Advances to Transform Histopathology Using Virtual Staining. BME Frontiers. https://doi.org/10.34133/2020/9647163.

Rodriguez, A. et al. (2014). Clustering by fast search and find of density peaks. Science. https://doi.org./10.1126/science.1242072.

Romero, J. M.. et al. (2020). Re-absorption and scattering of chlorophyll fluorescence in canopies: A revised approach. Remote Sensing of Environment, 246, 111860.

https://doi.org/10.1016/j.rse.2020.111860.

Rossini, M. et al. (2012). Remote sensing-based estimation of gross primary production in a subalpine grassland. Biogeosciences. https://doi.org/10.5194/bg-9-2565-2012.

Roth S. et al. (1978). Second harmonic generation in collagen. J. Chem. Phys. https://doi.org/10.1063/1.3132062.

Ruifrok, A. et al. (2001). Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology. PMID: 11531144.

Salah E. et al. (2021). A Fourier transform based audio watermarking algorithm. Applied Acoustic. https://doi.org/10.1016/j.apacoust.2020.107652

Saravanan, G. et al. (2016). Real time implementation of RGB to HSV/HSI/HSL and its reverse color space models. 2016 International Conference on Communication and Signal Processing. https://doi.org/ 10.1109/ICCSP.2016.7754179.

Schoukens, J. et al. (1986). Modeling the noise influence on the Fourier coefficients after a discrete Fourier transform. IEEE Transactions on Instrumentation and Measurement. https://doi.org/ 10.1109/TIM.1986.6499210

Scodellaro R. et al. (2019). Whole-section tumor micro-architecture analysis by a two-dimensional phasor-based approach applied to polarization-dependent second harmonic imaging. Front. Oncol. https://doi.org/10.3389/fonc.2019.00527.

Scodellaro R. et al. (2022). A novel hybrid machine learning phasor-based approach to retrieve a full set of solar-induced fluorescence metrics and biophysical parameters. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2022.113196.

Shaban, MT. et al. (2019). Staingan: Stain style transfer for digital histological images. IEEE 16th international symposium on biomedical imaging. https://doi.org/10.48550/arXiv.1804.01601

Siegmann, B. et al. (2019). The high-performance airborne imaging spectrometer HyPlant-from raw images to top-of-canopy reflectance and fluorescence products: Introduction of an automatized processing chain. Remote Sensing. https://doi.org/10.3390/rs11232760.

Tiaho, F. et al. (2007). Estimation of helical angles of myosin and collagen by second harmonic generation imaging microscopy. Opt. Express. https://doi.org/10.1364/OE.15.012286.

Tomczak, K. et al. (2015). The Cancer Genome Atlas (TGCA) : in immeasurable source of knowledge. Contemporary oncology. https://doi.org/10.5114/wo.2014.47136

Tsafas, V. et al. (2020) Polarization-dependent second-harmonic generation for collagen-based differentiation of breast cancer samples. Journal of Biophotonics. https://doi.org/10.1002/jbio.202000180

Vahadane, A. et al. (2016). Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging. https://doi.org/ 10.1109/TMI.2016.2529665

Vallmitjana, A. et al. (2021). Phasor-based image segmentation: machine learning clustering techniques. Biomedical Optics Express. https://doi.org/10.1364/BOE.422766

Van Der Tol, C. et al. (2009). An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. Biogeosciences, https://doi.org/10.5194/bg-6-3109-2009.

Van Der Tol, C. et al. (2014). Models of fluorescence and photosynthesis for interpreting measurements of solar-induced chlorophyll fluorescence. Biogeosciences, https://doi.org/10.1002/2014JG002713.

Van Der Tol, C. et al. (2019). The scattering and re-absorption of red and near-infrared chlorophyll fluorescence in the models Fluspect and SCOPE. Remote Sensing of Environment.

https://doi.org/10.1016/j.rse.2019.111292.

Van Rijthoven, M. et al. (2020). HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Medical Image Analysis. https://doi.org/10.48550/arXiv.2006.12230

Van Seijen M. et al. (2019). Impact of delayed and prolonged fixation on the evaluation of immunohistochemical staining on lung carcinoma resection specimen. Virchows Archiv. https://doi.org/10.1007/s00428-019-02595-9

Verhoef, W. et al. (2018). Hyperspectral radiative transfer modeling to explore the combined retrieval of biophysical parameters and canopy fluorescence from FLEX – Sentinel-3 tandem mission multi-sensor data. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.08.006.

Verrelst, J. et al. (2015). Global sensitivity analysis of the SCOPE model: What drives simulated canopy-leaving sun-induced fluorescence?. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2015.06.002.

Verrelst, Z. et al. (2019). Quantifying vegetation biophysical variables from imaging spectroscopy data: A review on retrieval methods. Surveys in Geophysics 40, 589 629. https://doi.org/10.1007/s10712-018-9478-y.

Verrelst, J. et al. (2021). Mapping landscape canopy nitrogen content from space using PRISMA data. ISPRS Journal of Photogrammetry and Remote Sensing. https://doi.org/10.1016/j.isprsjprs.2021.06.017.

Vu, QD. et al. (2019) "Methods for segmentation and classification of digital microscopy tissue images." Frontiers in bioengineering and biotechnology. https://doi.org/ 10.3389/fbioe.2019.00053

Xydias, D. et al. (2021). Three-dimensional characterization of collagen remodeling in cell-seeded collagen scaffolds via polarization second harmonic generation. Biomedical Optics Express. https://doi.org/10.1364/BOE.411501

Yang, X. et al. (2015). Solar-induced chlorophyll fluorescence that correlates with canopy photosynthesis on diurnal and seasonal scales in a temperate deciduous forest. Geophysical Research Letters. https://doi.org/10.1002/2015GL063201.

Yang YG. et al. (2014). Quantum cryptographic algorithm for color images using quantum Fourier transform and double random-phase encoding. Information Sciences. https://doi.org/10.1016/j.ins.2014.02.124.

Yang, H. et al. (2017). Chlorophyll fluorescence tracks seasonal variations of photosynthesis from leaf to canopy in a temperate forest. Global Change Biology. https://doi.org/10.1111/gcb.13590.

Wagniere G. (1986). Theoretical investigation of Kleinman symmetry in molecules. Applied Physics. https://doi.org/10.1007/BF00693918

Wang, Q. et al. (2015). Second-harmonic generation microscopy for assessment of mesenchymal stem cell-seeded acellular dermal matrix in wound-healing. Biomaterials. https://doi.org/ 10.1016/j.biomaterials.2015.03.011

Wang, Z. et al. (2016). A practical guide to The Cancer Genome Atlas (TCGA). Methods in molecular biology. https://doi.org/10.1007/978-1-4939-3578-9_6

Weber, G. (1981). Resolution of the fluorescence lifetimes in a heterogeneous system by phase and modulation measurements. J. Phys. Chem. https://doi.org/10.1021/j150608a006.

Wolf, J. (1983). Redundancy, the Discrete Fourier Transform, and Impulse Noise Cancellation. IEEE Transactions on Communications. https://doi.org/10.1109/TCOM.1983.1095820.

Wu, S. et al. (2018). Quantitative evaluation of redox ratio and collagen characteristics during breast

cancer chemotherapy using two-photon intrinsic imaging Biomedical opt express. https://doi.org/10.1364/BOE.9.001375

Zhao, F. et al. (2018). Reconstruction of the full spectrum of solar-induced chlorophyll fluorescence: Intercomparison study for a novel method. Remote sensing of environment. https://doi.org/10.1016/j.rse.2018.10.021

Acknowledgements

List of published peer-reviewed Articles

- R.Scodellaro, M.Bouzin, F.Mingozzi, F.Granucci, L.D'Alfonso, M.Collini, G.Chirico, L.Sironi "Collagen micro-architecture investigation in tumor sections by means of second harmonic generation signal multiphasor analysis coupled with non-supervised machine learning techniques." Il Nuovo Cimento C (2021). doi:10.1393/ncc/i2021-21139-9

- M.Bouzin, A.Zeynali, M.Marini, L.Sironi, R.Scodellaro, L.D'Alfonso, M.Collini, G.Chirico, "Multiphoton laser fabrication of hybrid-photo-activable biomaterials". Sensors (2021). doi:10.3390/s21175891

- M.Marini, M.Bouzin, R.Scodellaro, L.D'Alfonso, L.Sironi, F.Granucci, F.Mingozzi, G.Chirico, M.Collini "Quantitative active super-resolution thermal imaging: The melanoma case study". Biomolecular Concepts (2022). doi: 10.1515/bmc-2022-0015.

- R.Scodellaro, I.Cesana, L.D'Alfonso, M.Bouzin, M.Collini, G.Chirico, R.Colombo, F.Miglietta, M.Celesti, D.Schuettemeyer, S.Cogliati, L.Sironi "A novel hybrid machine learning phasor-based approach to retrieve a full set of solar-induced fluorescence metrics and biophysical parameters". Remote Sensing of Environment (2022). doi:10.1016/j.rse.2022.113196

List of Conference Contributions

Posters and oral contributions that I provided to national and international conferences during my PhD studies are reported below. Oral contributions have been underlined:

- R.Scodellaro, M.Bouzin, F.Mingozzi, F.Granucci, L.D'Alfonso, M.Collini, G.Chirico, L.Sironi "Collagen micro-architecture investigation in tumor sections by means of second harmonic generation signal multiphasor analysis coupled with non-supervised machine learning techniques." at SIF 2020.

- R.Scodellaro, M.Bouzin, F.Mingozzi, D.Panzeri, L.D'Alfonso, F.Granucci, M.Collini, G.Chirico, L.Sironi "AI-based second harmonic generation signal multiphasor analysis for collagen micro-architecture investigation in tumor sections" at M2L School 2021.

- R.Scodellaro, M.Bouzin, V.Remori, D.Panzeri, E.Giampieri, L.Marongiu, F.Mingozzi, L.D'Alfonso, M.Collini, F.Granucci, G.Chirico, L.Sironi "Phasor-based approach enhanced by supervised machine learning techniques for collagen micro-architecture characterization and automatic image segmentation" at BPS 2021.

- R.Scodellaro, D.Panzeri, D.Inverso, M.Bouzin, L.D'Alfonso, M.Collini, G.Chirico, L.Sironi "AI-based virtual staining and semantic segmentation of anatomopathology sections" at EBSA 2021.

- R.Scodellaro, I.Cesana, L,D'Alfonso, M.Bouzin, M.Collini, G.Chirico, R.Colombo, F.Miglietta, M.Celesti, D.Schuettemeyer, S.Cogliati, L.Sironi "AI-based spectral windows phasor approach to retrieve solar-induced fluorescence spectra at top-of-canopy and photosystem level" at SIF 2021.

- R.Scodellaro, D.Panzeri, D.Inverso, M.Bouzin, L.D'Alfonso, M.Collini, G.Chirico, L.Sironi "AI-based pipelines for the automated recognition of hepatocellular carcinoma and the semantic segmentation of virtually stained liver biopsies" at BPS 2022.

- R.Scodellaro, I.Cesana, L.D'Alfonso, M.Bouzin, M.Collini, G.Chirico, R.Colombo, F.Miglietta, M.Celesti, D.Schuettemeyer, S.Cogliati, L.Sironi "A novel hybrid machine

learning phasor-based approach to retrieve a full set of solar-induced fluorescence metrics and biophysical parameters from model simulations and field data" at LPS 2022.

- G.Traversa, R.Scodellaro, D.Fugazza, L.Sironi, M.Frezotti "Comparison of feature-tracking methods for glacier-velocity evaluation in Polar and Alpine regions" at LPS 2022.

- R.Scodellaro, D.Panzeri, E.Pagani, L.Tuzzi, L.D'Alfonso, M.Bouzin, M.Collini, G.Chirico, D.Inverso, L.Sironi "A new phasor-based method for a fast and accurate staining normalization of whole-slide biopsies" at PhD Students Meeting 2022.

- D.Panzeri, R.Scodellaro, C.Lancellotti, G.Chirico, L.Di Tommaso, L.Sironi "Computer aided iron quantification in digital images of liver biopsies" at PhD Students Meeting 2022.

- L.Sironi, R.Scodellaro, D.Panzeri, E.Pagani, L.Tuzzi, M.Bouzin, L.D'Alfonso, M.Collini, G.Chirico, D.Inverso "Virtual H&E staining and semantic segmentation of liver biopsies by means of a convolutional neural network coupled to a phasor-based algorithm" at SIF 2022.

- R.Scodellaro, I.Cesana, L.D'Alfonso, M.Bouzin, M.Collini, G.Chirico, R.Colombo, F.Miglietta, M.Celesti, D.Schuettemeyer, S.Cogliati, L.Sironi "i-φ-MaLe: a novel hybrid machine learning phasor-based approach to retrieve a full-set of solar-induced fluorescence metrics and biopbysical parameters" at SIF 2022.

- D.Panzeri, R.Scodellaro, G.Chirico, C.Lancellotti, L.Di Tommaso, L.Sironi "Computer aided iron quantification in digital images of liver biopsies" at SIF 2022.

- R.Scodellaro, D.Panzeri, E.Pagani, L.D'Alfonso, M.Bouzin, M.Collini, G.Chirico, D.Inverso, L.Sironi "A novel machine learning pipeline to analyze unstained liver biopsies and automatically quantify tumor-related structures" at ECP 2022.

- D.Panzeri, R.Scodellaro, G.Chirico, C.Lancellotti, L.Di Tommaso, L.Sironi "Computer aided iron quantification on liver biopsy whole-slide images" at ECP 2022.

Awards

- Winner of full scholarship for PhD Student (Biophysics - University of Milano-Bicocca).

- Winner of Best Communications Award during the SIF 2020 Conference (14-18 September 2020). My communication has been published in a special edition of "Il Nuovo Cimento – Colloquia and communications in Physics".

- Winner of full scholarship for the participation to the Biophysical Society Annual Meeting 2021 (Virtual, 5-10 March 2021), awarded by the Italian Society of Biophysics (SIBPA).

- Winner of full scholarship for the participation to the InnEO Summer School (Brasov, 18-24 July 2021).

- Winner of EBSA Travel Award 2021, awarded by the European Biophysical Societes' Association (Wien, 24-28 July 2021).

- Winner of BPS Travel Award 2022, awarded by the Biophysical Society during the 66° Annual Meeting in San Francisco, 19-23 February 2022.

- Selected for the Best Poster Session (IT in Pathology) during the 34° European Congress of Pathology (ECP 2022).

- Winner of Best Communications Award during the SIF 2022 Conference (12-16 September 2022). My communication will be published in a special edition of "Il Nuovo Cimento – Colloquia and communications in Physics".