Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2024, **16**, 2707

Statistical approaches to Raman imaging: principal component score mapping†

Elia Marin, *[abcd] Davide Redolfi Bristol, [aef] Alfredo Rondinella, [c] Alex Lanzutti[c] and Pietro Riello [f]

In this research, Raman imaging was employed to map various samples, and the resulting data were analyzed using a suite of automated tools to extract critical information, including intensity and signal-to-noise ratio. The acquired spectra were further processed to identify similarities and investigate patterns using principal component analysis. The objective of this study was to establish guidelines for investigating Raman imaging results, particularly when dealing with large datasets comprising thousands of relatively low-intensity spectra. The overall quality of the results was assessed, and representative locations were determined based on the main Raman bands. While automated software solutions are insufficient for removing baselines and fitting the data, statistical analysis proved to be a powerful tool for extracting valuable information directly from the raw spectral data. This approach enables the extraction of as much information as possible from large arrays of spectral data, even in complex cases where automated software may fall short. The findings of this study contribute to enhancing the analysis and interpretation of Raman imaging results, providing researchers with a robust methodology for extracting meaningful insights from complex datasets, reducing the amount of effort required during data interpretation and analysis.

Received 29th January 2024
Accepted 10th April 2024

DOI: 10.1039/d4ay00171k

rsc.li/methods

1. Introduction

Raman imaging is a powerful analytical technique that combines the principles of Raman spectroscopy and imaging to provide chemical maps of the surfaces of various samples.[1,2] Raman spectroscopy, named after its discoverer C. V. Raman, involves the inelastic scattering of photons by molecular vibrations, providing valuable information about molecular composition and structure.[3] By incorporating imaging capabilities into Raman spectroscopy, Raman imaging enables the acquisition of spatially resolved chemical information, allowing researchers to visualize and understand the distribution of different chemical species across a sample surface.[1,4,5]

Raman imaging offers several distinct advantages over traditional Raman spectroscopy, which primarily provides average information for the entire sample, or very limited areas. One significant advantage is, as the name implies, the ability to generate chemical maps.[6–8] Raman imaging can provide chemical maps of large portions of the sample, with a resolution that can reach about 1 μm. This enables researchers to visualize the distribution of different chemical compounds across the surface.[9] This capability allows for the identification of spatially varying components, including impurities,[10] contaminants,[11] and heterogeneous mixtures.[12] Thanks to this feature, in recent years, the researcher's attention has been moved also towards the use of Raman imaging in biomedical field to differentiate cells,[13] to visualize the responses to toxic compounds[14] and to diagnose the cancer state.[15]

Indeed one of the main advantage of Raman imaging, which is also shared by Raman spectroscopy in general, is its non-destructive nature. It requires minimal sample preparation and allows for the investigation of samples in their native state. This is particularly useful for biological samples, where conventional preparation procedures such as fixing and staining irreversibly alter the chemical composition and the biological properties.[16,17]

As Raman imaging provides detailed, localized molecular information, such as molecular vibrations, rotational states, and chemical bonding, it enables not just the identification but also the visualization of specific compounds, polymorphs, or

[a]Ceramic Physics Laboratory, Kyoto Institute of Technology, Sakyo-ku, Matsugasaki, 606-8585 Kyoto, Japan. E-mail: elia-marin@kit.ac.jp
[b]Department of Dental Medicine, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kamigyo-ku, Kyoto 602-8566, Japan
[c]Department Polytechnic of Engineering and Architecture, University of Udine, 33100, Udine, Italy
[d]Biomedical Research Center, Kyoto Institute of Technology, Sakyo-ku, Matsugasaki, Kyoto 606-8585, Japan
[e]Department of Immunology, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kamigyo-ku, Kyoto 602-8566, Japan
[f]Department of Molecular Science and Nanosystems, Ca' Foscari University of Venice, Via Torino 155, 30172 Venice, Italy

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ay00171k

phases present in the sample. Moreover, differences in noise level, background signals and fluorescence also contribute to the visual output of regions with different morphological, chemical and biological properties, even if the specific molecular vibrations cannot be properly identified.[18–20] The ability to obtain high resolution molecular-level information imaging makes Raman imaging a valuable tool in various scientific fields, including materials science,[21,22] chemistry,[23,24] pharmaceuticals,[25,26] and life sciences,[27,28] but Raman imaging also has some inherent limitations that must be considered when interpreting the obtained results. One disadvantage is the reduction in spectral resolution and intensity compared to conventional Raman spectroscopy. Due to larger number of spectra to be acquired in order to obtain spatial-resolved information, Raman imaging may exhibit reduced spectral resolution and lower intensity-over-noise ratios.[29,30]

In large-area mapping using Raman imaging, additional challenges arise that can impact the quality and accuracy of the results. Difficulties in maintaining precise focus throughout the entire imaging process may arise, as focusing a large area uniformly can be challenging. Additionally, there is a possibility of drifting during long acquisition times, leading to misalignment or distortion in the resulting chemical maps, or the samples can undergo physical and chemical changes over time, for example due to water evaporation or oxidation.

When Raman maps are successfully acquired, the operator will obtain thousands if not hundreds of thousands of relatively weak and noisy raw Raman spectra. The data analysis will then typically progress along one of two potential pathways.

(1) If the typical Raman spectrum for that type of sample is known, the operator will extract maps representing the relative intensity of each of the known main bands, trying to emphasize their distribution across the map.

(2) If the typical Raman spectrum for that type of sample is unknown, before being able to extract valuable information, the operator will be forced to investigate various locations, trying to manually identify local variations in relative intensity.

In both cases, various statistical methods can be used to simplify and automatize the process of extracting and identifying crucial data through Raman imaging, without relying on the ability of the operator to manually screen the data to observe trends and anomalies. Various acquisition and post-treatment commercial software produce automated visual outputs,[31–34] but they usually don't provide the user with sufficient statistical information to understand the reliability and significance of the methods, resulting in visually appealing Raman maps in which the contributions of noise, fluorescence and surface morphology cannot be easily discriminated.

When comparing Raman imaging with similar techniques, it becomes evident that each technique has its strengths and limitations. Electron backscatter diffraction (EBSD) provides crystallographic information but does not provide chemical information with the same level of specificity as Raman imaging.[30] Energy-dispersive X-ray spectroscopy (EDS) allows for elemental analysis but lacks the ability to differentiate between different molecular species.[35,36] Fourier transform infrared spectroscopy (FTIR) imaging,[37] on the other hand, utilizes infrared radiation to probe molecular vibrations but has lower spatial resolution compared to Raman imaging and is sensitive to water vapor.

In this paper, we will focus on two different methods of information extraction that do not require data preprocessing (cosmic ray filter, baseline removal, smoothing, *etc.*) and can help operators identify sub-groups of spectra with similar characteristics, that can potentially maximize the amount of information extracted from Raman imaging data. After statistical analysis, the sub-groups can then be post-analyzed by conventional means, labeled and used to extract additional information such as Raman shift, band intensity and band full-width at half maximum.

Similar combinations of Raman imaging and PCA analysis have been previously successfully utilized, in particular for much smaller datasets with very high signal over noise ratios,[38] where the algorithm was capable to distinguish between three phases and also extract fitting parameters. In the case of tens of thousands of complex Raman spectra consisting of tens of bands each, such an approach would be impracticable as it would be impossible to process without the use of a supercomputer.

In other cases, the combination of Raman imaging and PCA analysis, again on very small datasets,[39] required heavy mathematical pre-treatment in order to be able to extract relevant information.[40]

When large datasets have been utilized (tens of thousands of spectra), literature results were focused on a more theoretical approach, and did not provide chemical maps showing the distribution of the different phases,[41] or were limited to very specific scenarios.[42] Nevertheless, these studies were successful in demonstrating that the combination of Raman imaging and PCA analysis can be utilized to perform analysis of large portions of sample surfaces, within reasonable time frames.

For comparison, our analytical method is completely automated and universal. It can process datasets containing tens or even hundreds of thousands of spectra of any size in a reasonable time without requiring any pre-treatment of the dataset. Additionally, it provides information not only about phase presence and distribution but also about the quality of the dataset itself. Moreover, to the best of the authors' knowledge, this manuscript presents the first performance comparison of Raman imaging combined with PCA for samples of different natures.

The six samples used in this manuscript were selected to cover various scenarios that scientists working in the field of material science might encounter when using Raman spectroscopy. No biological samples or liquids were included, as they require additional care in sample preparation and are often not stable over time. An explanation of how to extend these methods to biological samples will be published in the near future. Calibrations and testing, performed on artificial Raman datasets, can be found in the ESI.†

The same statistical tools can be directly translated to other similar spectroscopic techniques commonly used to obtain compositional "maps" of the sample surface, such as

cathodoluminescence, electron backscatter diffraction or Auger and energy dispersive X-ray spectrometry.

2. Experimental

2.1 Sample preparation

In this work, six different samples were analyzed using Raman spectrometry and Raman imaging. The sample typology, chemical composition and morphology are resumed in Table 1 and Fig. 1.

The rationale behind the six samples is to try to cover various potential real life scenarios, focusing on materials science. The blank sample is utilized to verify that the methods don't produce artifact maps when the Raman spectra are extremely noisy. The reference sample is used to confirm that the methods don't create more sub-regions than necessary, giving the user the false impression that there are different phases present, the ginkgo leaf shows the potential of this methodology for very broad and strongly overlapping Raman bands, the ZrO_2/PE has a clear interface and two clearly distinguishable phases so that the results of the PC analysis can be compared with the optical images, the AlN/PMMA composite has a dispersion of secondary phase inside a matrix, with also a relatively high surface roughness, which constitutes a more challenging scenario for this type of analyses, and the sintered Si_3N_4 is a ceramic well-known for its network of sub-micrometric secondary phases at the grain boundaries, that go often undetected by conventional Raman spectroscopy.

2.2 Raman imaging

Raman imaging maps were obtained using a dedicated Raman device (RAMANtouch, Nanophoton Co., Minoo, Osaka, Japan) operated in microscopic measurement mode with confocal imaging capability in two dimensions. Making use of a linear detector, this model of Raman microscope can achieve simultaneous image acquisition of 400 spectra. It used an excitation source of 532 nm. All maps were acquired at 5× magnification, with an excitation power density of about 200 W cm^{-2} and a 300 g mm^{-1} grating reaching a spectral resolution of about ~2 cm^{-1}. The CCD utilized for these experiments was a Pixis Excelon 400 (Teledyne Princeton Instruments, Quakerbridge, New Jersey, United States) with a CCD format of 1340 × 400, 20 × 20 μm pixels operating with a 100% fill factor. Maps were acquired using a dedicated software (RAMANview, Nanophoton Co., Minoo, Osaka, Japan), then exported as *.txt files with the Raman shift as the first column and the spatial coordinates as the first row. At 5× magnification, the lateral resolution of the maps was 4.15 μm (Table 2).

2.3 Statistical tools

In these six examples of applications, the *.txt data arrays were processed using Python, in this sequence.

(1) Raw spectra loading: the code loads the Raman spectra from the raw data skipping the first raw where the spatial coordinates of the spectra are stored. The first column, where the Raman shifts are stored, is saved separately, but not used for the data processing.

(2) Normalization: the spectra are normalized calling the function "StandardScaler()" from the scikit-learn library (sklearn) is a preprocessing technique used to standardize features in a dataset. Standardization involves transforming the features in a way that they have a mean of 0 and a standard deviation of 1. This is particularly useful in machine learning algorithms that rely on the scale of features, as it helps to ensure that all features are on a similar scale, preventing certain features from dominating others due to their larger magnitudes. "StandardScaler()" calculates the mean and standard deviation of each feature in the training data. Then, it transforms the data by subtracting the mean from each feature and dividing by the standard deviation. This process ensures that the transformed data has a standard normal distribution (mean of 0 and standard deviation of 1).

(3) Clusterization: the arrays were split in clusters by similarities. In these practical examples the number of clusters was usually set to 3, but it can be lower or higher depending on the complexity of the map itself and the level of detail the code is expected to be able to discriminate. The clustering was performed using the "K-means" algorithm from the scikit-learn library. The "K-means" algorithm aims to minimize the within-cluster sum of squares, also known as the inertia or distortion. It seeks to find cluster centroids that minimize the distance between data points and their assigned centroid, effectively creating compact and well-separated clusters. After convergence, the "K-means" algorithm provides the final cluster centroids and assigns each data point to a specific cluster. The resulting clusters can be analyzed to gain insights into the underlying patterns or groupings within the data. The main drawback of using the "K-means" algorithm is its sensitivity to the initial selection of centroids, which can result in variations in the clustering process output.

Table 1 Description of the six different samples

Sample	Description	Composition	Shape
Blank	Map acquired with the laser off	N/A	N/A
Si reference	Acquired on a silicon wafer	Si (100)	25.4 mm × 1 mm wafer
Ginkgo	*Ginkgo biloba* leaf, fresh (picked in summer)	Lignin, cellulose, *etc.*	Natural shape, uncut, 1 hour after collection
ZrO_2/PE	Interface between zirconia and polyethylene	ZrO_2, $(C_2H_4)_n$	5 mm × 3 mm × 20 mm plates, pressed together
AlN/PMMA	15 wt% AlN reinforced PMMA, 3D printed by stereolithography	AlN, $(C_5O_2H_8)_n$	5 mm × 30 mm × 30 mm plate
Si_3N_4	Sintered Si_3N_4	β-Si_3N_4	10 mm × 20 mm × 20 mm plate

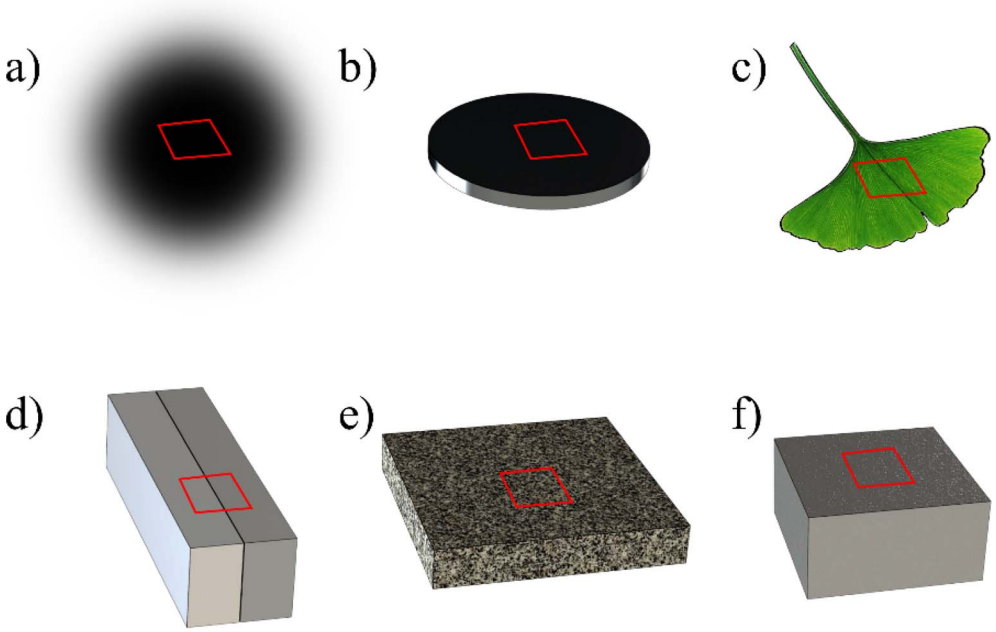

Fig. 1 Drawings representing the six different samples used in this research: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

(4) Grouping: spectra are then grouped based on the cluster labels created in (3).

(5) Calculation of the Signal Ratio (SR): the intensity ratio is computed as the ratio of the length of the curve (considered to be a simple model for measuring intensity while limiting the influence of the background signal) to the total number of data points in the spectrum. It provides an indication of the relative intensity or strength of the Raman signals in each spectrum. The general equation is presented in eqn (1):

$$\mathrm{SR}_{x,y} = \frac{\sum_{i=1}^{n} |I_i - I_{i-1}|}{n} \quad (1)$$

where x and y are the coordinates of the point, n is the length of the array for the Raman shift and I is the spectral intensity at the position i of the Raman shift array. Benchmark testing on the Signal Ratio can be found in the ESI.†

(6) Calculation of Signal-to-Noise Ratio (SNR): SNR ratios can be calculated using various methods, depending on the structure of the data and the desired output. In this protocol, the SNR ratio was calculated as the previously calculated SR and the same ratio in a limited Raman shift range, which extends from 2000 cm^{-1} to 2200 cm^{-1}, which is considered to be a "silent zone" in most Raman applications. As the Raman signal in the "silent zone" is approximately zero, the only contribution is produced by the noise and the SNR will be higher for points that have bigger differences in specific length between the whole spectrum and the silent zone. The SNR is ultimately calculated using eqn (2):

$$\mathrm{SNR}_{x,y} = \frac{\mathrm{SR}_{x,y}}{N_{x,y}} = \frac{\dfrac{\sum_{i=1}^{n} |I_i - I_{i-1}|}{n}}{\dfrac{\sum_{k=k_i}^{k_f} |I_k - I_{k-1}|}{k_f - k_i}} \quad (2)$$

where N is the noise, k_i is the starting index of the noise region (the closest point over 2000 cm^{-1}), in these examples, and k_f is the final index of the noise region (the closest point over 2200 cm^{-1}). Benchmark testing on the Signal-to-Noise Ratio can be found in the ESI.†

(7) Average spectra: for each cluster, the code calculates the average spectra, removes the baseline with an automated

Table 2 Acquisition parameters for the 6 different samples

Sample	Exposure time	Lines	Data points	Map height	Time
Blank	1 × 10 s	150	60 000	620 μm	42 min
Si reference	1 × 10 s	150	60 000	620 μm	42 min
Ginkgo	2 × 1 s	200	80 000	970 μm	7 min
ZrO_2/PE	1 × 15 s	100	40 000	500 μm	3h 42 min
AlN/PMMA	3 × 30 s	147	58 800	600 μm	3h 42 min
Si_3N_4	1 × 300 s	100	40 000	600 μm	8h 17 min

procedure and smooths the curve. This step is purely used as a feedback for the operator, as the pre-processed data is not stored or utilized for statistical analysis. The operator can compare the average spectra of the clusters, spot errors and estimate if the number of cluster selected (c) is adequate for the specific application.

(8) Principal Component Analysis (PCA): PCA is performed on the normalized spectra data using the "PCA()" function from the scikit-learn library. PCA is a dimensionality reduction technique commonly used in machine learning and data analysis. PCA aims to transform a dataset consisting of potentially correlated variables into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they explain in the original data. The main purpose of PCA is to reduce the number of dimensions (features) in the dataset while retaining as much information as possible. This can help in various ways, such as speeding up computation, improving model performance, and visualizing high-dimensional data. The principal components are linear combinations of the original features, where the first principal component explains the maximum variance in the data, the second principal component explains the second most variance, and so on. By choosing a smaller number of principal components, you can capture a significant portion of the variance while reducing the dimensionality of the data. For the purpose of this paper, the dimensional parameter is arbitrarily set to 3. The scores represent the transformed data in the reduced-dimensional space. Scatter plots for PC1 *vs.* PC2, PC1 *vs.* PC3 and PC2 *vs.* PC3 are then generated, using the cluster labels to visually discriminate between the groups created in (4).

(9) Principal components weighted maps: while PCA scatter plots show the scores of single Raman spectra with respect to the principal components in the reduced dimensional space, plot maps of the PC scores at each location where Raman spectra were collected can show the scores for each principal component as a function of their coordinates. Each map represents the distribution of scores across the entire sample's surface. Higher (or lower) scores in a specific PC indicate that the corresponding feature is prominent (or less prominent) at that location. Analyze the PC score maps to gain insights into the sample's composition, structure, and variations. Features in the score maps can help identify different components within the sample and understand how they vary spatially. Benchmark testing for the influence of parameters such as background intensity, cosmic ray density, noise and signal intensity on the PC maps can be found in the ESI.†

3. Results

Fig. 2 shows the surface morphology of the different samples, as observed through the 5× magnification lens used for Raman imaging. As Fig. 2(a) was obtained in the absence of any real sample inside the microscope chamber, the resulting image is completely black. For Fig. 2(b), the gray surface of the silicon wafer appears to be uniform, and no features could be identified at low magnifications. In Fig. 2(c), the morphology of the ginkgo leaf and in particular the distribution of the cells and three of the principal veins, running from the top to the bottom of the image, are clearly visible even if blurred due to the limited depth of field of the microscope. In Fig. 2(d), the white material on the left is composed of 3YSZ (tetragonal zirconia stabilized by a 3% of yttria), while the darker material on the right, characterized by oriented marks due to machining, is ultra-high molecular weight polyethylene. In between the two bars, a 100 micron gap is clearly visible. On Fig. 2(e) the dispersion of AlN particles in the PMMA matrix can be appreciated even at low magnifications, and the oriented marks created on the surface by the path of the laser during the stereolithographic process are also clearly visible, going from the top left to the bottom right of the image. Fig. 2(f) share some similarities with Fig. 2(b), but it also features a dispersion of barely visible darker dots, located at the grain boundaries of the silicon nitride crystals.

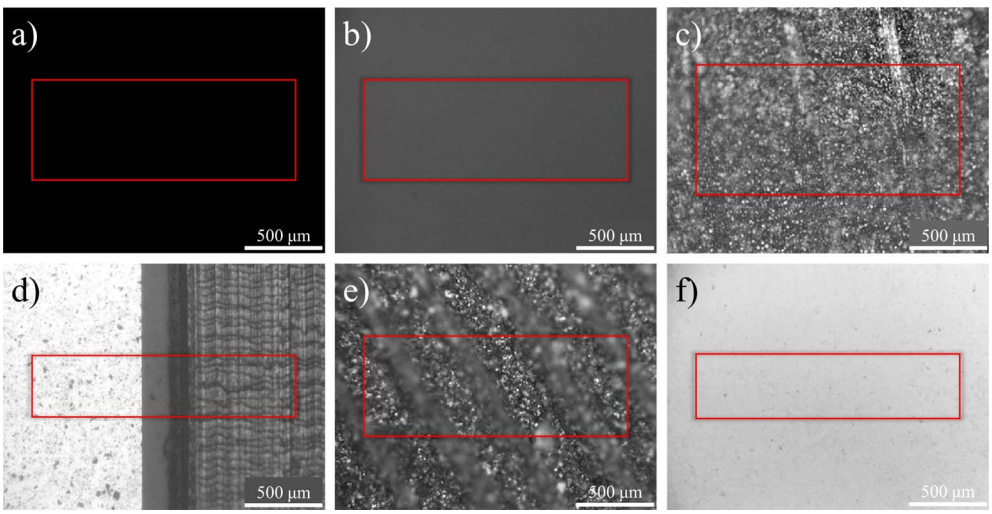

Fig. 2 Optical images of the six different samples as observed through the same 5× objective lens used for Raman imaging: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

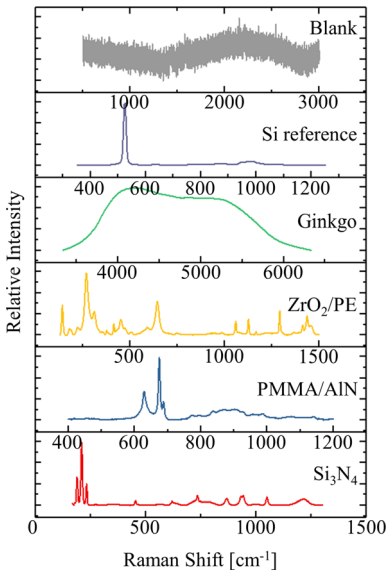

Fig. 3 Average Raman spectrum for each of the investigated areas on the six different samples. The raw average spectra (without pre-processing) are presented in the ESI, Fig. S21.†

Fig. 3 shows the average Raman spectrum acquired on each of the six different samples, in a region considered to be representative and after baseline removal. It can be observed that the spectra of the Blank sample are purely constituted by noise, as expected, while the Si reference is dominated by the intense band at 520 cm^{-1} and associated to crystalline Si. The ginkgo leaf shows strong fluorescence in the region between 3500 and 6000 cm^{-1} due to the presence of chlorophyll, the ZrO_2/PE sample features bands related to vibrational modes of both materials (ZrO_2 in the region up to 700 cm^{-1}, followed by PE between 1000 and 1500 cm^{-1}), the PMMA/AlN map appears to be mainly dominated by the strong bands of AlN, between 600 and 700 cm^{-1}, while the bands of PMMA, between 700 and 1200 cm^{-1}, appear to be much weaker. Ultimately, the Si_3N_4 average spectrum features all the band associated with this material in previous literature. A complete list of the band positions associated with each of the phases in Fig. 3 can be found in the ESI,† as it goes beyond the scope of this article.

In order to properly estimate the quality of the Raman imaging maps, the code was implemented with an algorithm to evaluate the intensity of the Raman signal, as described in Section 2.3, step 4.

The maps in Fig. 4 don't provide detailed information about the nature of the signal detected or the quality of the signal itself, as noise and background signals also contribute to the modulus of $I_i - I_{i-1}$ in eqn (1). Two of the samples, in Fig. 4(a) and (b), have their intensity color bars starting from values higher than zero, despite both maps being basically railed at the bottom of the scale. This might indicate that in both cases the signal is relatively low in intensity and the contribute of the noise significant. Fig. 4(c), on the other hand, shows two distinct regions, with the veins and some regions of the leaf showing a stronger signal than others. By the nature of eqn (1), local changes in the intensity of the fluorescence signal would reflect in the SR values, for example as a consequence of variations in the local chlorophyll content. On the map taken at the interface between ZrO_2 and polyethylene of Fig. 4(d), the ZrO_2 region on the left has a generally higher SR when compared to the polyethylene on the right, and this is in line with the differences of relative intensity between the bands on the left and right side of Fig. 3, associated to ZrO_2 and polyethylene,

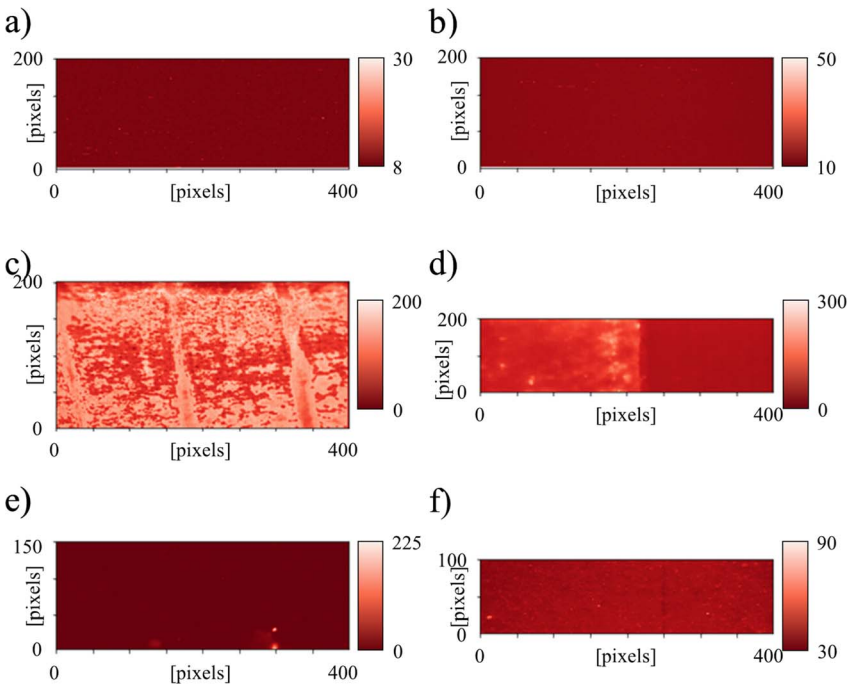

Fig. 4 SR maps as obtained on the six different samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

respectively. Fig. 4(e), related to the composite material containing PMMA and AlN, is somehow similar to Fig. 4(a) and (b), but two regions with high SR values can be observed on the bottom of the image. Those spikes do not correspond to topographical anomalies in Fig. 2, and can be caused by either different phases or more intense background and/or noise signals. The last map, related to a synthesized Si_3N_4 block (Fig. 4(f)), features a dispersion of small areas with relatively intense Raman signal on an otherwise "dark" matrix, possibly due to intergranular phases.

Instead of focusing purely on the intensity of the signal, Fig. 5 shows the distribution of the signal-to-noise ratio, SNR, as defined in Section 2. Due to the high heterogeneity between spectra from different materials, differences in relative intensity, background signals, fluorescence peak shape and peak number, it is difficult to univocally define noise and various algorithms can be used to estimate the signal-to-noise ratio, leading to discording and even contradictory results. For the sake of these working examples, noise has been defined as the data scattering in the region between 2000 and 2200 cm^{-1}, which is often considered a "silent zone" in Raman spectroscopy due to the absence of characteristic peaks, with the notable exceptions of C≡C and C≡N.

Despite many similarities, the maps in Fig. 4 and 5 have a different scope and provide complementary information. In Fig. 5(a) for the Blank sample and Fig. 5(b) for the Si reference, the signal is not just low, but also comparable to the noise, meaning that the exposure time of the latter was relatively low when compared to the other maps, even if the average spectrum in Fig. 3 appears to be sufficiently intense. The vibrational information of other possible phases, with Raman intensity lower than the silicon reference, is likely to have been lost. Fig. 5(c) on the other hand perfectly replicates Fig. 4(c), and this is caused by the extremely high intensity of the fluorescence signal when compared to the background noise. In contrast, Fig. 5(d) shows an inversion in relative intensity between the two phases, with the SNR of polyethylene higher than the SNR of ZrO_2 despite the SR of ZrO_2 being higher than the SR of polyethylene. This effect is caused by the different nature of the two materials, their uniformity, the level of fluorescence and the way their physical structure interacts with the laser beam. In Fig. 5(e), the relatively intense regions observed in Fig. 4(e) are barely visible, meaning that despite the increase in intensity they maintain the SNR more or less constant. The last panel, Fig. 5(f) has been obtained with very long exposure times of about 5 minutes for each line. The long exposition increased the chances for cosmic rays to hit the detector and generate very intense but extremely localized but completely random spikes in the Raman signal. The effect of the spikes is barely visible in the SR map of Fig. 4(f), but their relative intensity is enhanced by the structure of eqn (2).

Fig. 6 shows the distribution of the three groups of spectra on the maps, as defined by their similarity following the results of the K-means algorithm from the scikit-learn library. The algorithm is fast and simple, but it is not particularly sensitive to singularities such as Raman peaks. Spectra are usually grouped together depending on their "shape", but other factors such as relative intensity and background signal can also play a role. In Fig. 6(a) obtained without turning on the laser, for example, all spectra are completely random and the algorithm ended up trying to find similarities in the electrical noise in the background, resulting in a regular grid of colors. On the Si reference map of Fig. 6(b), on the other hand, all spectra are morphologically similar, but the relative intensity is higher in

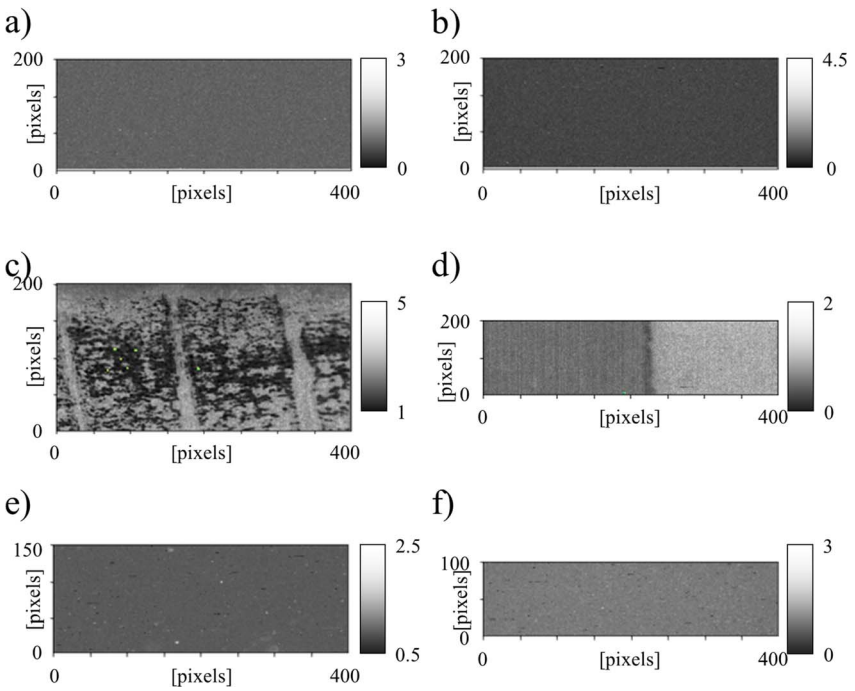

Fig. 5 SNR maps as obtained on the six different samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO2/PE, (e) AlN/PMMA, (f) Si_3N_4.

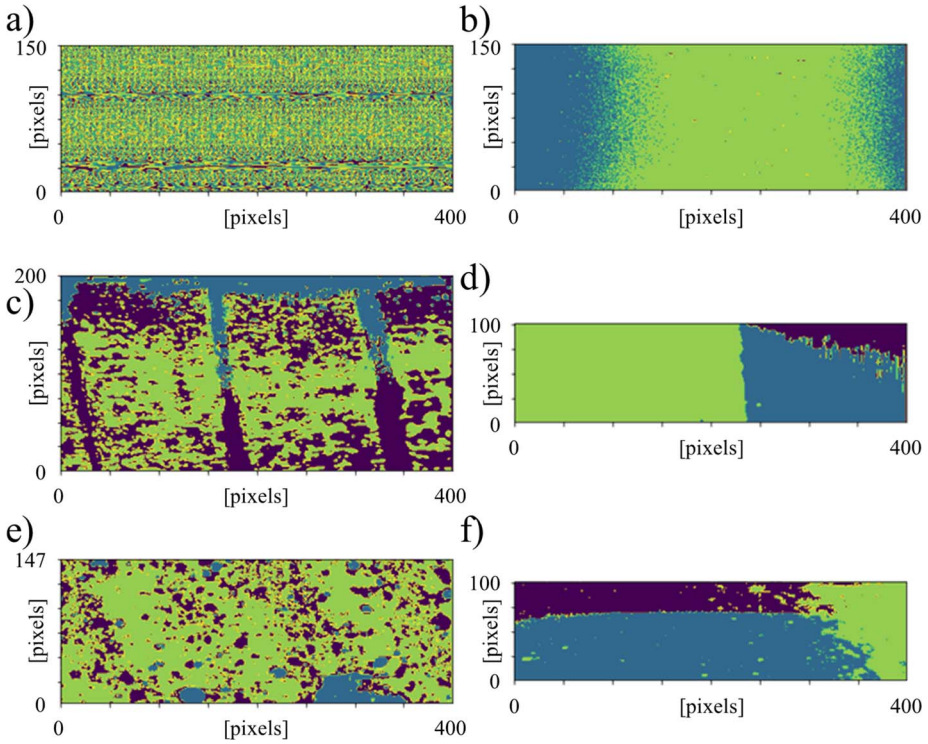

Fig. 6 3-Cluster maps of the six different samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

the central region because of the lens spherical aberration, resulting in two different groups of spectra divided at an "arbitrary" point, as the algorithm is unable to understand that there is a more or less continuous gradient of intensity. Fig. 6(c) shows some interesting similarities with both the SR and the SNR maps of Fig. 4(c) and 5(c), and while the differences between the blue and purple regions resulted to be caused by relative intensity, green areas also featured a different ratio between the two broad peaks related to chlorophyll fluorescence and located at about 4200 cm^{-1} and 5200 cm^{-1}. The two materials of Fig. 6(d), on the other hand were similarly split into three groups because of a misalignment of the polyethylene side, where the upper part of the image resulted to be closer to the optical lens than the bottom, resulting in differences in the intensity. If two groups were to be used instead of three, the relative grouping would be solely based on the shape of the spectra. In Fig. 6(e), related to the composite material containing AlN particles embedded into a PMMA matrix, the algorithm was able to discriminate the presence of particles (purple in figure), but only in the regions where the sample surface was perfectly on focus. Due to the nature of the printing process, some regions of the surface are protruding more than others (scan lines), and the depth of focus of the lens was not sufficient to compensate. In the blue regions, no Raman signal other than fluorescence could be measured, which explains the differences between the SR and the SNR maps of Fig. 4(e) and 5(e). In Fig. 6(f), the three regions are a consequence of both the tilted surface (higher on the top, purple, and lower on the bottom, blue) and the lens aberration (higher on the right, green).

It should be noted that in these working examples the number of clusters has been arbitrarily set to 3, but the number should be adjusted depending on the complexity of the array, the quality of the spectra, the homogeneity within each phase present, the morphology of the surface and the scope of the analysis. A calibration of the number of clusters for the AlN/PMMA and ZrO_2/PE samples is presented in the ESI (Fig. S22).†

Fig. 7 shows the results of the PCA analysis performed on the six samples, and in particular the correlation between the PC1 and PC2 scores, while the colors represent the clusters of Fig. 6. As expected, the scores related to the PC of the Blank samples are very low (Fig. 7(a)), as the spectra are almost completely random. On the Si reference sample (Fig. 7(b)), on the other hand, due to the differences in relative intensity caused by the spherical aberration, most points lay on a diagonal line but again with low PC1 and PC2 scores. For the ginkgo leaf sample, the scattering of the points closely resembles both Fig. 7(a) and (b), but with much higher scores. In this case PC2 could be associated with the relative intensity of the spectra, while PC1 resulted to be correlated to the relative intensity ratio between the two main bands. In the case of the ZrO_2/PE sample (Fig. 7(d)), the high score PC1 resulted to be correlated to the ratio between the zirconia bands and the polyethylene bands, while the much lower score PC2 was correlated to the relative intensity of the spectra. In the case of the AlN/PMMA composite (Fig. 7(e)), the scores for both PC1 and PC2 are relatively low, and this is caused by various factors: the fraction of AlN particulate inside the PMMA matrix, the large Raman scattering cross-section of AlN, the transparency of PMMA to the laser and the roughness of the surface. For all these reasons, the Raman

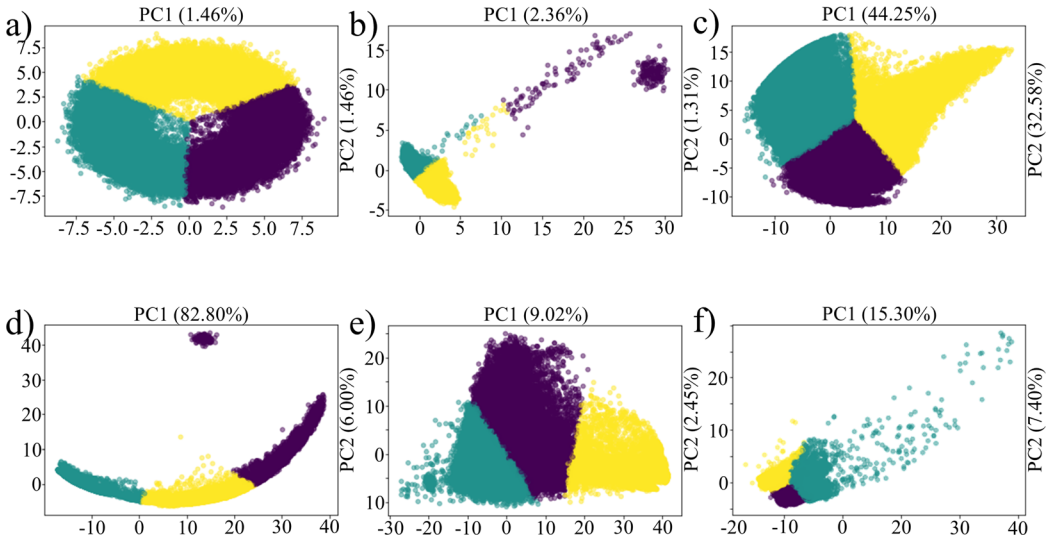

Fig. 7 PC1 and PC2 scores for the six different samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

spectra collected at all points appear to be all a linear combination between a relatively strong AlN and a relatively week PMMA signals, despite the fraction of PMMA being more than 4 times higher. Variations in the scores of PC1 represent different ratios between the two phases, while PC2 scores are related to the overall intensity of the spectra. For the last sample, the sintered Si_3N_4 block of Fig. 7(f), the scatter appears to be comparable to that of Fig. 7(b), but with higher scores for both PC1 and PC2. In this case, PC1 scores are mainly determined by the presence of an inter granular phase where the three main bands of Si_3N_4 are shifted and weak secondary peaks appear. PC2, on the other hand, is mainly contributed by the intensity of the signal, which is weaker on the inter granular phases, resulting in an almost linear correlation between PC1 and PC2 in those regions. Scores scatter plots for PC2 *vs.* PC3 and PC1 *vs.* PC3 can be found in the ESI (Fig. S23 and S24).†

On the maps of Fig. 8, the scores assigned to the first component for each point of the maps is represented with

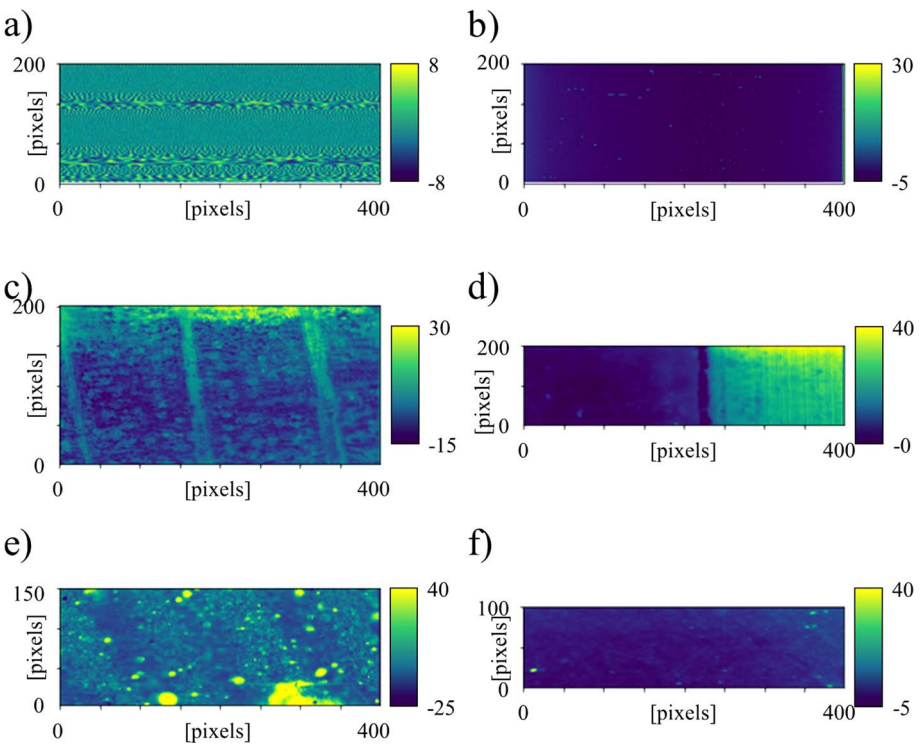

Fig. 8 Shows a map of the scores of the first principal component (PC1) as a function of location, for each of the six samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

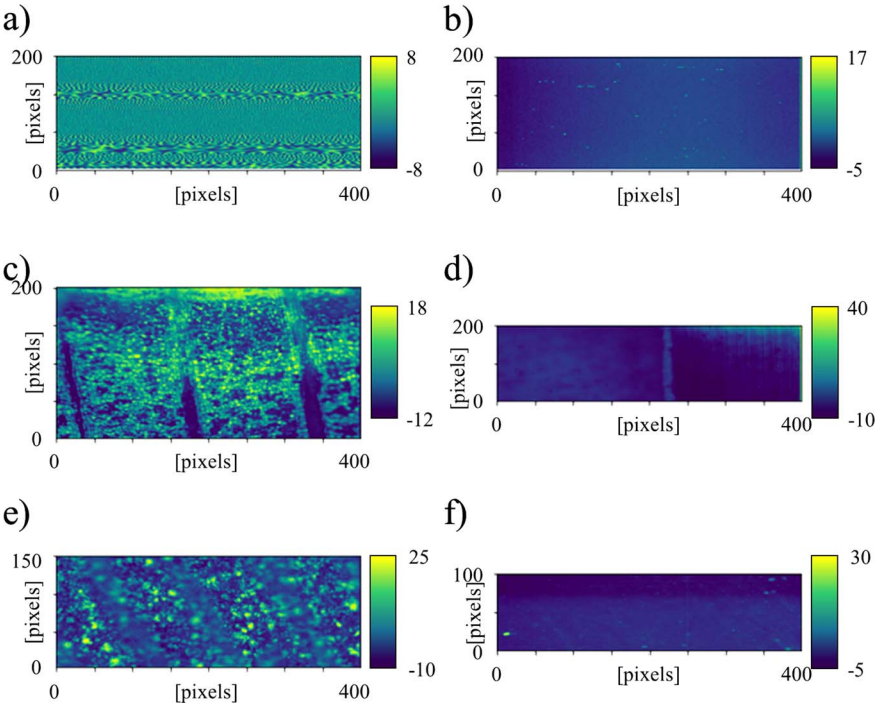

Fig. 9 Shows a map of the scores of the second principal component (PC2) as a function of location, for each of the six samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

a color scale. While the map in Fig. 8(a), related to the Blank sample, is clearly influenced by the presence of noise, Fig. 8(b) and (f) seem to provide equally little information, as points with high PC1 scores are present, but randomly distributed. This is further evident when the scores from PC2 and PC3 are plotted, as in Fig. 9 and 10: for samples that are almost completely uniform in distribution, such as the Si reference and the Si_3N_4 block, what can be determined by PC is the presence of spectroscopic defects or the location of contaminants. Secondary phases also contribute to the PC scores, but as observed for the Si_3N_4 block their relative intensity is so small that they often get confused with other defects and contaminations. Despite the lack of features for the average spectra of the gingko leaf in Fig. 3, PC maps do provide useful information about the composition, distribution and morphology of the sample. For PC1, in Fig. 8, a stronger score is detected on the veins when compared to the rest of the leaf, and on the top of the image when compared to the bottom. This distribution is mainly based on the intensity of the band at 4100 cm^{-1} and seem to correlate well with the total amount of chlorophyll, where veins detain higher levels of chlorophyll when compared to other parts of the leaf with veins closer to the axil having the highest concentration. When the scores of the PC2 are mapped, the presence and fraction of oxidized chlorophyll (fluorescence band centered at about 5250 cm^{-1}) can be detected as part of the second principal component (Fig. 9(c)), with cluster of neighboring cells either showing relatively high or relatively low scores. The small region with high third principal component scores see their Raman spectra degenerated, with the secondary peak shifted from 5250 cm^{-1} to about 5000 cm^{-1}, and are characterized by an overall lower emission. For the ZrO_2/PE sample, the map of the first principal component gives stronger scores for the polyethylene bar when compared to zirconia, with the gap between the two having the lowest score. In this sample, each phase has higher scores in a separate principal component, with the three phases being polyethylene (PC1, Fig. 8(d)), gap region (PC2, Fig. 9(d)) and zirconia (PC3, Fig. 10(d)), with PC3 accounting for only the 1.5% of the variance because of the high scattering and intense background noise. In the case of the composite AlN/PMMA material, the first principal component in Fig. 8(e) has an large variance despite being caused by fluorescence and not proper Raman scattering. In the most intense regions of the map, the Raman signal is completely lost, covered by the strong background signal. The second principal component, mapped in Fig. 9(e), on the other hand, perfectly follows the distribution of the AlN particulate inside the PMMA matrix. The third principal component in Fig. 10(e), which showed relatively low scores at specific, round locations, could be associated with residual sub-superficial micro-bubbles that did not appear during the morphological analysis of Fig. 2.

4. Discussion

During the statistical analysis of large Raman imaging datasets, various approaches can be used in order to be able to extract meaningful information. In this manuscript, we showed a few examples of application of one of these possible approaches, intended for a broad range of potential applications. In these analysis, the datasets have not been pre-treated or filtered in order to reduce noise or to perform a baseline removal, as these

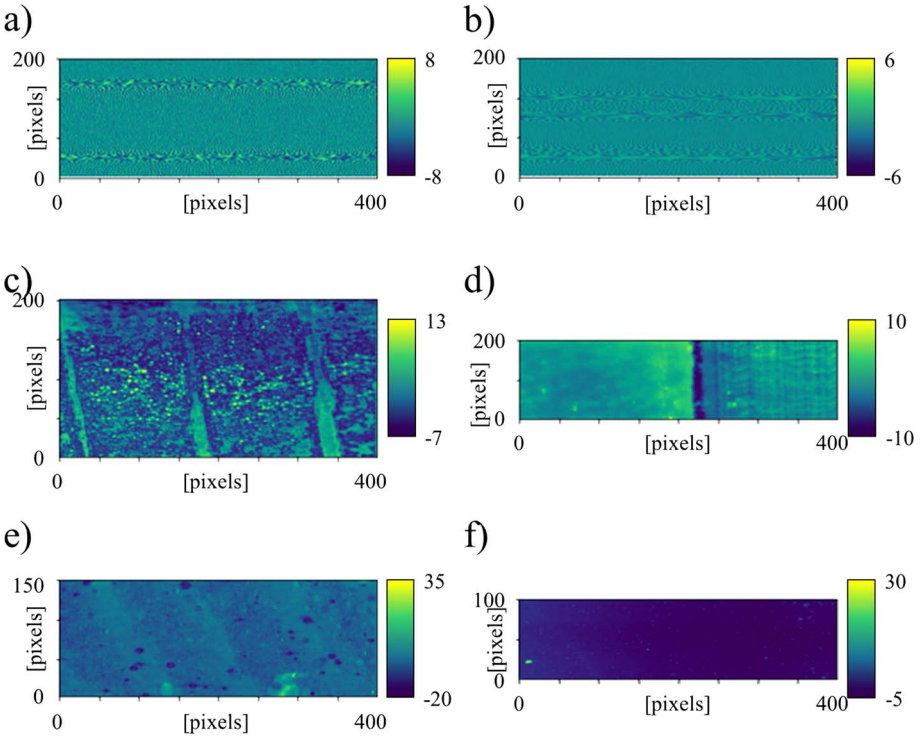

Fig. 10 Shows a map of the scores of the third principal component (PC3) as a function of location, for each of the six samples: (a) blank, (b) Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4.

operations can hardly be automated without the risk of losing potentially useful information. On large datasets with thousands of Raman spectra, even user-assisted data pre-treatment is not applicable unless the whole dataset is sufficiently uniform so that baseline removal and filtering can be performed based on a limited knowledge of the contents of the database itself.

Performing data analysis on the average spectra, such as those presented in Fig. 3, lead to the loss of individual spectral information, oversimplification of otherwise heterogeneous samples, increased influence of background signals and inadequate representation of minor components, such as secondary phases. To mitigate these limitations and risks, it is generally advisable to consider analyzing both the average spectrum and individual spectra within the dataset, but this process, which is often performed in Raman mapping, gets progressively more time consuming with increasing the number of spectra.

The processes proposed in this manuscript are not intended as a substitute for data analysis and processing, but they provide potentially useful information to reduce the burden on the operator by grouping together Raman spectra with similar characteristics and giving information on the overall quality of the investigated area. Moreover, different algorithms can be used to achieve similar results, as the focus is on the sequence of operations that can provide the user with the largest amount of useful information in the shortest time possible.

By measuring the Signal Ratio (SR) of each acquired spectra, the user can immediately estimate the amount of spectral information that can be extracted at different locations of the samples. Where the SR is relatively low with respect to the rest of the map, the region is more likely to have less Raman scattering signal, more noise or higher background signals. Low SR values does not automatically mean that no useful information can be extracted by those regions, but maps such as the ones provided in Fig. 4 simplify the initial screening processes when determining the quality of the results. For Fig. 4(a), for example, we could safely assume than any randomly picked point would be representative for the whole investigated area.

In Fig. 5, the SNR is simply calculated as the total length of the curve divided by the number of spectral points, but such as simple algorithm alone can't discriminate between signal and noise. For this reason, the maps of the Blank sample and the Si reference are comparable and uniform in color.

By observing the six SR maps of Fig. 4 alone, the user can expect that in four samples, the Blank, the Si reference, the AlN/PMMA and the Si_3N_4 block, the spectra all share similar amounts of spectroscopic information, meaning that the average spectra in Fig. 2 are probably well representative for the whole surface.

By performing an additional SNR analysis (Fig. 5), the user can visualize the areas that have the strongest signal with respect to the background noise. Depending on the specific application, different definitions of noise can be applied, and to measure noise in Raman Spectroscopy is further complicated by the large variability in peak intensity, peak width and number of peaks. For the sake of these working examples, the noise has been defined as the previously defined SR compared to the length of the curve in a specific region of the Raman spectrum, often called "silent zone". The results of Fig. 4 and 5 are similar for very uniform samples, such as the Blank and the Si

reference, but provide complementary information when the region is not homogeneous. In the case of the ginkgo leaf, for example, the central region of the veins gives a low SR signal because those regions are out of focus, so the lens it's not able to collect back the same amount of scattered light. Despite the lower intensity, their SNR intensity is similar to that of the surrounding regions, meaning that the spectra have comparable quality. For the ZrO_2/PE sample, SR and SNR maps are complementary, meaning that the zirconia have higher signal intensity but also higher background noise, making the SNR actually better for the less intense polyethylene region. On the AlN/PMMA composite and the Si_3N_4 block, the SR maps show a few locations of higher intensity due to fluorescence, while the few dots on the SNR maps are caused by cosmic rays hitting the CCD camera.

In these working example, a simple clusterization algorithm has been used to split the maps in regions of "similarity", based on the Euclidian distance from the centroids. In the first step, the algorithm begins by randomly initializing K (with $K = 3$) cluster centroids in the feature space. These centroids act as the representative points for the clusters. Each spectrum is then assigned to the nearest centroid based on the Euclidean distance. There are other ways to address "distance" (or similarity) other than Euclidean, such as correlation distance, spectral angle distance or spectral information divergence. Euclidean distance has been chosen due to its simplicity, broad applicability and fast implementation, but in most real case scenarios the results would benefit from the use of more elaborated methods.

It can be observed that the output of a clusterization based on the Euclidean distance of a spectrum from the K centroids, is quite different from the maps of Fig. 4 and 5, which are based on relative intensity at each spectral interval. As a result, clustermaps seem to be sensitive to differences in chemical composition (reflected into the presence/absence of specific peaks at specific locations, such as in Fig. 6(d)), but also surface alignment, as clearly shown in Fig. 6(b). By choosing the right number of clusters (K), it is possible to obtain a combination of clues of both morphological and topological origin. The map obtained on the AlN/PMMA composite for $K = 6$ in Fig. S22 in the ESI,† for example, is very similar to the micrograph presented in Fig. 2(e). To be able to extract the maximum amount of useful information from a simple clusterization would require the operator to progressively increase (or decrease) the value of K until the most "meaningful" result is obtained, with "meaningful" arbitrarily defined. From the chemical composition point of view, two clusters are sufficient in the case of the ZrO_2/PE interface and three in the case of the AlN/PMMA composite, but this evaluation is only possible after the average spectra of each cluster are compared.

After evaluating "how" the maps can be easily clusterized depending on the spectroscopic differences between different sub-regions, the next step would be to estimate how much similar (or different) the spectra are, in this case by using PCA. PCA is a technique that transforms the original features of the dataset into a new set of linearly uncorrelated variables called principal components. The number of components determines the dimensionality of the reduced dataset and the number can either be decided by the operator or it can be set so that the number of components that capture a certain amount of variance in the data are set automatically. The transformed data will have a reduced dimensionality, where each sample will be represented by the specified number of components. The principal components are obtained by taking linear combinations of the original features, with each component representing a different direction in the feature space.

By comparing the cluster map of Fig. 6 with the scatter plots of Fig. 7, the former appears to be a brutal but effective approximation of the latter, as each cluster of each map is well-differentiated and defined in the principal components score space. The first evidence of the robustness of the method is that the variances of the principal components are low for investigated areas that didn't show clear differences in the SR and SNR maps and had either scattered or alignment-dependent clusterizations.

Respect to the cluster maps, the PC score maps of Fig. 8–10 are less influenced by the topography of the surface. Phases, porosities, contaminations and defects are the main sources of spectroscopic differences detected by PC maps. Only for the AlN/PMMA composite sample, the high roughness caused by the printing process results in localized blurring.

Fig. 11 shows the average spectra of four samples, with the main contributions of each PC marked in blue, red and green respectively. The contributions could be easily calculated by comparing the average spectrum of the whole map with the spectra of the regions with high PC1, PC2 and PC3 scores, respectively.

For gingko, these analyses proved to be useful in the discrimination of chlorophyll and chlorophyll derivatives. High PC1 scores could be obtained from regions with strong chlorophyll-a fluoresce emissions, high PC2 scores for regions with strong chlorophyll-b fluorescence emissions and PC3 with pheophytin-a.

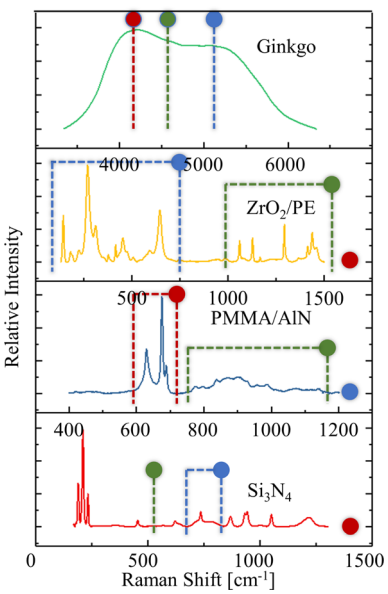

Fig. 11 Average spectra of Ginko, ZrO_2/PE, PMMA/AlN and Si_3N_4, with the regions associated high PC1 (blue), PC2 (red) and PC3 (green) scores.

In the case of the ZrO_2/PE interface, PC1 scores were correlated to the intensity of the zirconia region, PC2 scores with high noise regions (the gap between the two materials) and PC3 with polyethylene.

For the PMMA/AlN composite, due to high roughness the PC1 scores were influenced by noise (distance between lens and sample), PC2 could detect the AlN particulate and PC3 the polymeric matrix.

In the case of Si_3N_4, PC1 scores could be associated with intergranular phases, which generate a broad but not very intense band in the region between 700 and 800 cm^{-1}. Like in the case of ZrO_2/PE, the PC2 scores were high for areas with high noise, in particular small porosities. PC3 scores could be associated with a band at about 520 cm^{-1} and caused by unreacted silicon crystals.

In all four cases, the principal components score maps were able to extract useful spectroscopic information from the datasets, even if the noise caused by the topography of the surface (PMMA/AlN), the presence of gaps in the surface (ZrO_2/PE) or other defects (Si_3N_4) was often associated with one of the first three principal components. This information can be utilized by the user to simplify data extraction procedures and rapidly detect meaningful spectral differences in large datasets. Representative spectra extracted from locations with high PC-scores can be then utilized to perform conventional Raman analysis such as spectral deconvolution and labeling.

5. Conclusions

Simple statistical methods were successfully used to extract useful information from Raman imaging databases. The simple procedures proposed in this research could be used to determine the quality of datasets acquired during Raman imaging and identify meaningful differences in the Raman spectra, even in the presence of intense noise.

By utilizing clustering methods, the maps could be roughly divided in regions by spectral similarity, depending on their distance from the centroid.

More accurate results could be achieved using a new protocol of principal component score mapping, which allowed to not only identify the main spectroscopic components present in the dataset, but also their distribution across the surface of the samples.

By using a similar protocol, it would be possible to reliably extract the valuable information stored in large Raman imaging datasets, that at the present time require a lot of effort from the users.

Conflicts of interest

There are no conflicts to declare.

References

1 P. J. Treado and M. D. Morris, Infrared and Raman Spectroscopic Imaging, *Appl. Spectrosc. Rev.*, 1994, **29**(1), 1–38.
2 H. Kano, H. Segawa, P. Leproux and V. Couderc, Linear and nonlinear Raman microspectroscopy: History, instrumentation, and applications, *Opt. Rev.*, 2014, **21**(6), 752–761.
3 R. S. Krishnan and R. K. Shankar, Raman effect: History of the discovery, *J. Raman Spectrosc.*, 1981, **10**(1), 1–8.
4 S. Stewart, R. J. Priore, M. P. Nelson and P. J. Treado, Raman Imaging, *Annu. Rev. Anal. Chem.*, 2012, **5**(1), 337–360.
5 M. C. Caggiani and P. Colomban, Raman microspectroscopy for Cultural Heritage studies, *Phys. Sci. Rev.*, 2018, **3**(11)), available from, **https://www.degruyter.com/document/doi/10.1515/psr-2018-0007/html**.
6 K. E. Shafer-Peltier, A. S. Haka, J. T. Motz, M. Fitzmaurice, R. R. Dasari and M. S. Feld, Model-based biological Raman spectral imaging, *J. Cell. Biochem.*, 2002, **87**(S39), 125–137.
7 N. Gierlinger, L. Sapei and O. Paris, Insights into the chemical composition of Equisetum hyemale by high resolution Raman imaging, *Planta*, 2008, **227**(5), 969–980.
8 N. Anderson, P. Anger, A. Hartschuh and L. Novotny, Subsurface Raman Imaging with Nanoscale Resolution, *Nano Lett.*, 2006, **6**(4), 744–749.
9 J. L. Xu, K. V. Thomas, Z. Luo and A. A. Gowen, FTIR and Raman imaging for microplastics analysis: State of the art, challenges and prospects, *TrAC, Trends Anal. Chem.*, 2019, **119**, 115629.
10 F. A. Ponce, J. W. Steeds, C. D. Dyer and G. D. Pitt, Direct imaging of impurity-induced Raman scattering in GaN, *Appl. Phys. Lett.*, 1996, **69**(18), 2650–2652.
11 X. Xiao, X. Liu, T. Mei, M. Xu, Z. Lu, H. Dai, *et al.*, Estimation of contamination level in microplastic-exposed crayfish by laser confocal micro-Raman imaging, *Food Chem.*, 2022, **397**, 133844.
12 K. C. Polavaram and N. Garg, Enabling phase quantification of anhydrous cements *via* Raman imaging, *Cem. Concr. Res.*, 2021, **150**, 106592.
13 J. Gala De Pablo, M. Lindley, K. Hiramatsu and K. Goda, High-Throughput Raman Flow Cytometry and Beyond, *Acc. Chem. Res.*, 2021, **54**(9), 2132–2143.
14 K. Dodo, K. Fujita and M. Sodeoka, Raman Spectroscopy for Chemical Biology Research, *J. Am. Chem. Soc.*, 2022, **144**(43), 19651–19667.
15 L. Becker, N. Janssen, S. L. Layland, T. E. Muerdter, A. T. Nies, K. Schenke-Layland, *et al.*, Raman imaging and fluorescence lifetime imaging microscopy for diagnosis of cancer state and metabolic monitoring, *Cancers*, 2021, **13**(22), 5682.
16 Z. Movasaghi, S. Rehman and I. U. Rehman, Raman Spectroscopy of Biological Tissues, *Appl. Spectrosc. Rev.*, 2007, **42**(5), 493–541.
17 A. C. S. Talari, Z. Movasaghi, S. Rehman and I. U. Rehman, Raman Spectroscopy of Biological Tissues, *Appl. Spectrosc. Rev.*, 2015, **50**(1), 46–111.
18 D. Zhang and D. Ben-Amotz, Enhanced chemical classification of Raman images in the presence of strong fluorescence interference, *Appl. Spectrosc.*, 2000, **54**(9), 1379–1383.
19 H. Abramczyk, J. Surmacki, M. Kopeć, A. K. Olejnik, A. Kaufman-Szymczyk and K. Fabianowska-Majewska, Epigenetic changes in cancer by Raman imaging,

fluorescence imaging, AFM and scanning near-field optical microscopy (SNOM). Acetylation in normal and human cancer breast cells MCF10A, MCF7 and MDA-MB-231, *Analyst*, 2016, **141**(19), 5646–5658.

20 G. J. Puppels, T. B. Schut, N. M. Sijtsema, M. Grond, F. Maraboeuf, C. G. De Grauw, *et al.*, Development and application of Raman microspectroscopic and Raman imaging techniques for cell biological studies, *J. Mol. Struct.*, 1995, **347**, 477–483.

21 A. Cantarero, Raman scattering applied to materials science, *Procedia Mater. Sci.*, 2015, **9**, 113–122.

22 A. Rousaki and P. Vandenabeele, In situ Raman spectroscopy for cultural heritage studies, *J. Raman Spectrosc.*, 2021, **52**(12), 2178–2189.

23 R. L. McCreery, *Raman Spectroscopy for Chemical Analysis*, [Internet], John Wiley & Sons, 2005, [cited 2024 Mar 22], available from, **https://books.google.com/books?hl=en&lr=&id=qY4MI0Zln1YC&oi=fnd&pg=PR5&dq=23.%09McCreery+RL+Raman+Spectroscopy+for+Chemical+Analysis,+John+Wiley+%26+Sons,+2000&ots=nf_axT_Y_N&sig=q9m5YrGSfgjbGST-aecE9Doz6P8**.

24 H. Yamakoshi, K. Dodo, M. Okada, J. Ando, A. Palonpon, K. Fujita, *et al.*, Imaging of EdU, an Alkyne-Tagged Cell Proliferation Probe, by Raman Microscopy, *J. Am. Chem. Soc.*, 2011, **133**(16), 6102–6105.

25 S. Wartewig and R. H. Neubert, Pharmaceutical applications of Mid-IR and Raman spectroscopy, *Adv. Drug Delivery Rev.*, 2005, **57**(8), 1144–1170.

26 P. J. Gallimore, N. M. Davidson, M. Kalberer, F. D. Pope and A. D. Ward, 1064 nm Dispersive Raman Microspectroscopy and Optical Trapping of Pharmaceutical Aerosols, *Anal. Chem.*, 2018, **90**(15), 8838–8844.

27 I. Notingher and L. L. Hench, Raman microspectroscopy: a noninvasive tool for studies of individual living cells *in vitro*, *Expert Rev. Med. Devices*, 2006, **3**(2), 215–234.

28 K. A. Antonio and Z. D. Schultz, Advances in Biomedical Raman Microscopy, *Anal. Chem.*, 2014, **86**(1), 30–46.

29 M. Sattlecker, C. Bessant, J. Smith and N. Stone, Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics, *Analyst*, 2010, **135**(5), 895–901.

30 J. P. Rolland, P. Meemon, S. Murali, K. P. Thompson and K. S. Lee, Gabor domain optical coherence microscopy, in *Design and Quality for Biomedical Technologies III*, [Internet], SPIE, 2010, pp. 42–50, [cited 2024 Mar 22], available from, **https://www.spiedigitallibrary.org/conference-proceedings-of-spie/7556/75560A/Gabor-domain-optical-coherence-microscopy/10.1117/12.849009.short**.

31 T. Yamamoto, J. N. Taylor, S. Koseki and K. Koyama, Classification of food spoilage bacterial species and their sodium chloride, sodium acetate and glycine tolerance using chemometrics analysis and Raman spectroscopy, *J. Microbiol. Methods*, 2021, **190**, 106326.

32 T. Adachi, F. Boschetto, N. Miyamoto, T. Yamamoto, E. Marin, W. Zhu, *et al.*, In vivo regeneration of large bone defects by cross-linked porous hydrogel: a pilot study in mice combining micro tomography, histological analyses, Raman spectroscopy and synchrotron infrared imaging, *Materials*, 2020, **13**(19), 4275.

33 H. Imamura, W. Zhu, T. Adachi, N. Hiraishi, E. Marin, N. Miyamoto, *et al.*, Raman analyses of laser irradiation-induced microstructural variations in synthetic hydroxyapatite and human teeth, *J. Funct. Biomater.*, 2022, **13**(4), 200.

34 G. Pezzotti, M. Kobara, E. Marin, W. Zhu, I. Nishimura, O. Mazda, *et al.*, Raman imaging of pathogenic Candida auris: Visualization of structural characteristics and machine-learning identification, *Front. microbiol.*, 2021, **12**, 769597.

35 W. C. Campbell, Energy-dispersive X-ray emission analysis. A review, *Analyst*, 1979, **104**(1236), 177–195.

36 L. J. Allen, A. J. D'Alfonso, B. Freitag and D. O. Klenov, Chemical mapping at atomic resolution using energy-dispersive x-ray spectroscopy, *MRS Bull.*, 2012, **37**(1), 47–52.

37 S. E. Glassford, B. Byrne and S. G. Kazarian, Recent applications of ATR FTIR spectroscopy and imaging to proteins, *Biochim. Biophys. Acta, Proteins Proteomics*, 2013, **1834**(12), 2849–2858.

38 O. A. Maslova, G. Guimbretière, M. R. Ammar, L. Desgranges, C. Jégou, A. Canizarès, *et al.*, Raman imaging and principal component analysis-based data processing on uranium oxide ceramics, *Mater. Charact.*, 2017, **129**, 260–269.

39 H. Shinzawa, K. Awa, W. Kanematsu and Y. Ozaki, Multivariate data analysis for Raman spectroscopic imaging, *J. Raman Spectrosc.*, 2009, **40**(12), 1720–1725.

40 A. S. Blervacq, M. Moreau, A. Duputié, I. De Waele, L. Duponchel and S. Hawkins, Raman spectroscopy mapping of changes in the organization and relative quantities of cell wall polymers in bast fiber cell walls of flax plants exposed to gravitropic stress, *Front. Plant Sci.*, 2022, **13**, 976351, available from, **https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.976351/full**.

41 J. Hutchings, C. Kendall, B. Smith, N. Shepherd, H. Barr and N. Stone, The potential for histological screening using a combination of rapid Raman mapping and principal component analysis, *J. Biophot.*, 2009, **2**(1–2), 91–103.

42 C. Fang, Y. Luo, X. Zhang, H. Zhang, A. Nolan and R. Naidu, Identification and visualisation of microplastics *via* PCA to decode Raman spectrum matrix towards imaging, *Chemosphere*, 2022, **286**, 131736.

