Electronic Supplementary Material (ESI) for Analytical Methods.

Statistical approaches to Raman imaging: principal component score mapping

Elia Marin[a,b,c,d], Davide Redolfi Bristol[a,e,f], Alfredo Rondinella[c], Alex Lanzutti[c], Pietro Riello[f]

[a]Ceramic Physics Laboratory, Kyoto Institute of Technology, Sakyo-ku, Matsugasaki, 606-8585 Kyoto, Japan;
[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

1. Dummy Raman generator

To evaluate the efficiency of the analytical method, we created a dummy Raman generator capable of producing rectangular Raman arrays containing 400 x 200 spectra, with each spectra going from 400 to 4000 cm^{-1}. This was done in order to replicate the possible input of the analytical software, while being able to control key factors such as noise and background intensity. The dummy generator is able to replicate cosmic rays, background (parabolic or linear), noise, as well as 4 independent Raman signals with different distributions. For each component, the Raman spectrum and the distribution is presented in Figures S1 and S2. In brief:

- Signal 1(Fig. S1b and S2b) contains one sharp band at 500 cm^{-1} and a broad weak band at 1500 cm^{-1}, localized in vertical stripes;
- Signal 2 (Fig. S1c and S2c) contains three sharp bands, at 510 cm^{-1}, 1200 cm^{-1} and 2900 cm^{-1}, with different relative intensities and FWHM, and their absolute intensity is a gradient from the left to the right of the map;
- Signal 3(Fig. S1d and S2d) contains a sharp band located at 1750 cm^{-1}, and is distributed inside three circles of different diameter;
- Signal 4(Fig. S1e and S2e) is located inside four circles of different diameter, that partially overlap with both Signal 1 and Signal 3 distributions.

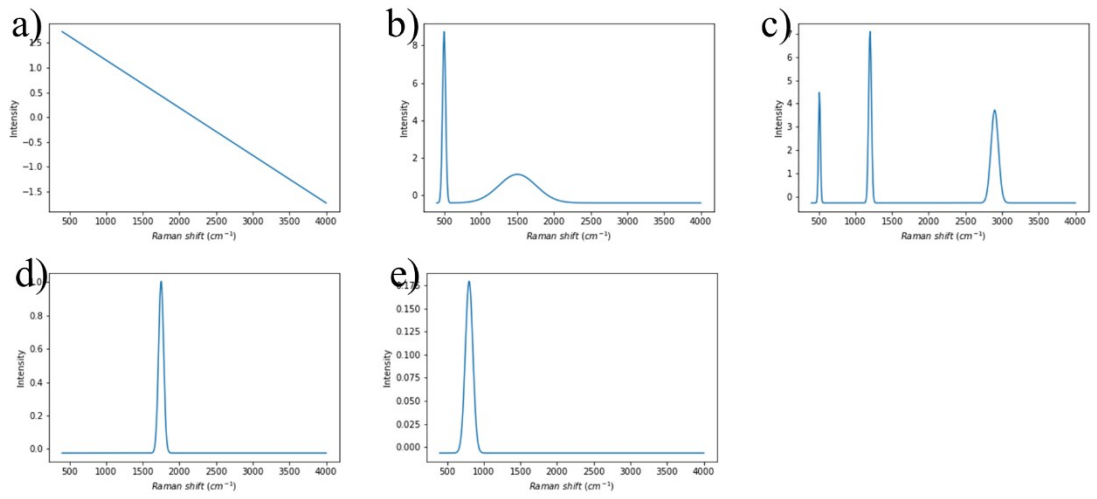

Figure S1: dummy Raman signals used by the software to produce simulated Raman datasets: a) a linear background (example), b) Signal 1, c) Signal 2, d) Signal 3 and e) Signal 4

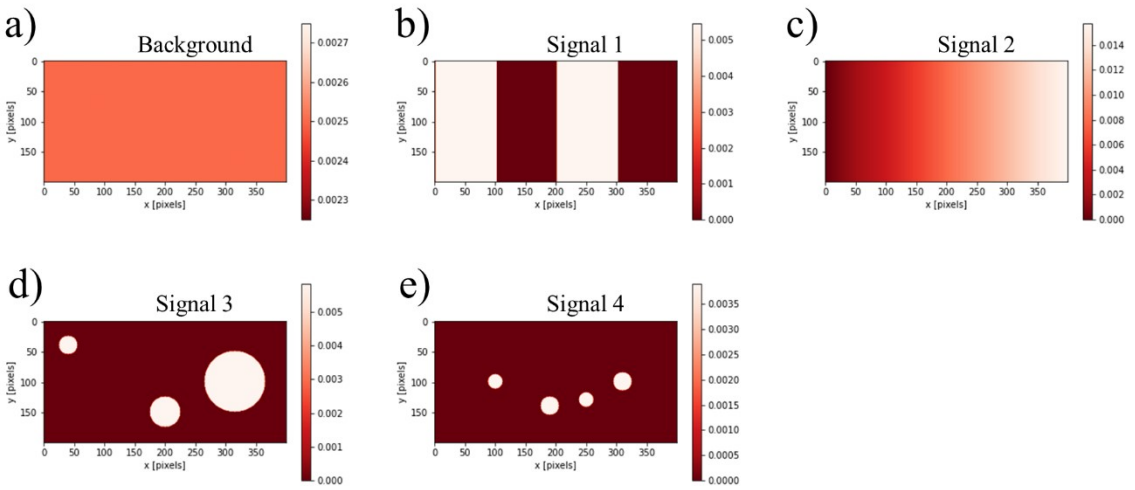

Figure S2: distribution of the different Raman signals on the simulated Raman datasets

2. Benchmarks

The analytical method analyzes Raman maps and performs Principal Component Analysis (PCA) to identify the presence of different chemical components within a sample. Here's how we define success based on the number of distinct signals visualized in the PCA analysis, using 3 PC:

- Complete Success (4 Signals): All four expected chemical components are clearly distinguishable and represented by separate signals in the PCA image. This indicates the software accurately identified and differentiated all components within the sample;
- Partial Success (3 Signals): Three out of the four expected chemical components are clearly visible in the PCA image. This suggests the software performed well in identifying most components, but one may be obscured by noise or by the other signals;
- Failure (2 or Less Signals): Two or fewer distinct signals are present in the PCA image. This indicates the software may not be effectively differentiating the different chemical components within the sample, at these simulating conditions.

3. Effect of cosmic rays

The effect of cosmic rays on the PCA analysis (for an otherwise empty Raman dataset) is summarized in Figure S3 and S4, as a function of increased probability, as summarized in Table S1:

Table 1: probability of having a cosmic ray for each data point and for each spectra

Image	Probability (for a single data point)	Probability (for a spectra)
S3a	0	0
S3b	10^{-5}	4×10^{-3}
S3c	10^{-4}	4×10^{-2}
S3d	5×10^{-4}	0.2
S3e	2.5×10^{-3}	1
S3f	1.25×10^{-2}	5

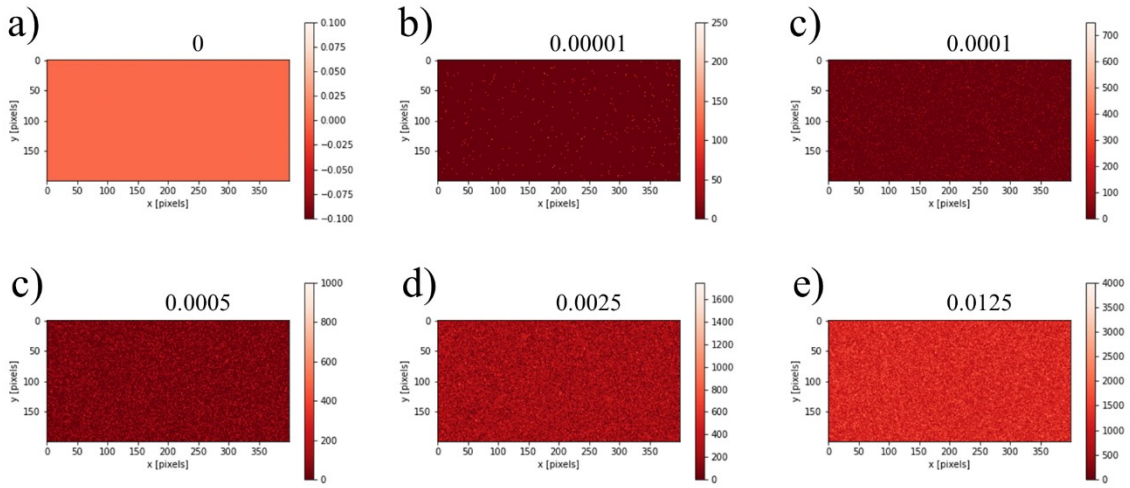

Figure S3: intensity maps as a function of the probability for a data point to be a cosmic ray: a) 0 b) 10^{-5} c) 10^{-4} d) 5×10^{-4} e) 2.5×10^{-3} f) 1.25×10^{-2}

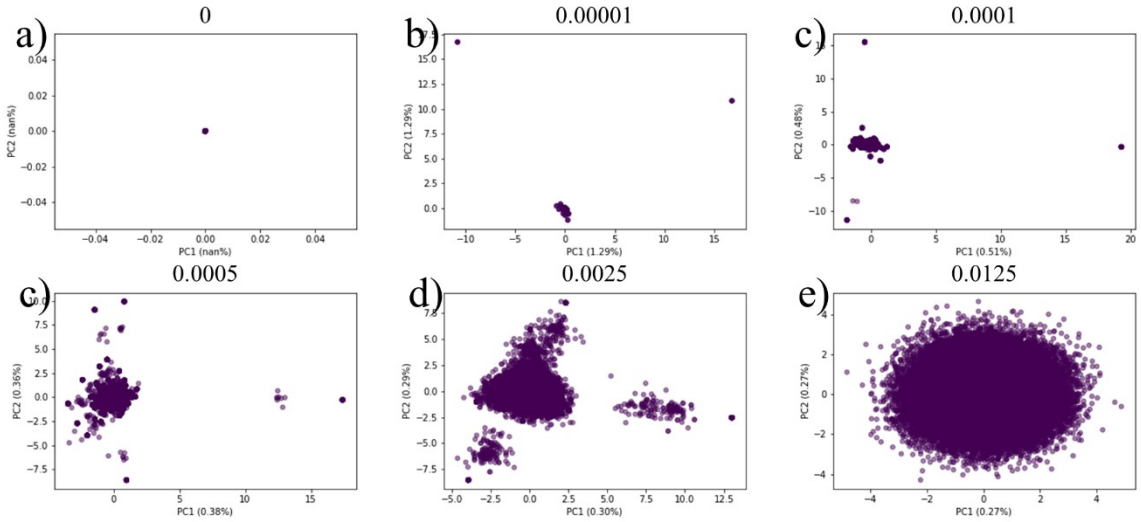

Figure S4: PC1 *vs* PC2 as a function of the probability for a data point to be a cosmic ray: a) 0 b) 10^{-5} c) 10^{-4} d) 5 x 10^{-4} e) 2.5 x 10^{-3} f) 1.25 x 10^{-2}

In Figure S4, some clustering can be observed already at cosmic ray probability 0.0005, but the PC1 and PC2 scores of these clusters appear to be very low, meaning that in presence of any other Raman signal distributions, these clusters are unlikely to be visible after a PC analysis.

4. Effects of noise

The effect of noise has been taken into account by adding a randomized signal proportional to the maximum intensity of the Raman signals. Examples with relative intensity between 0.1 and 10 are presented in Figure S5, S6, S7, S8 and S9, where, for simplicity, all four Raman signals were set at the same maximum intensity. The subfigure a) represent the intensity map (as described in the main manuscript), the subfigure b) the result of the clustering process (as described in the main manuscript), the subfigure c) the signal over noise ratio (as described in the main manuscript), the subfigures d) e) and f) the scores of the PC1, PC2 and PC3 respectively.

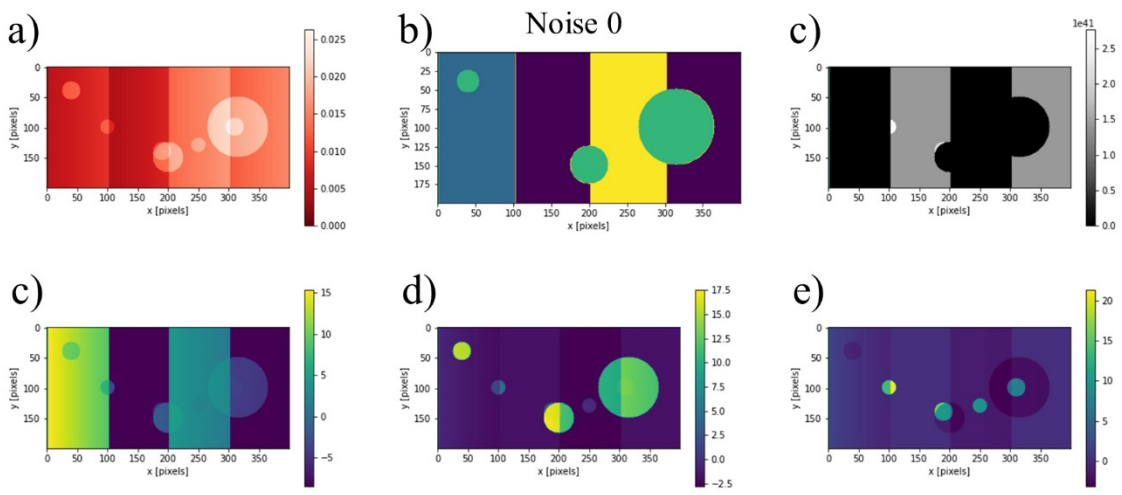

Figure S5: analysis performed in absence of noise

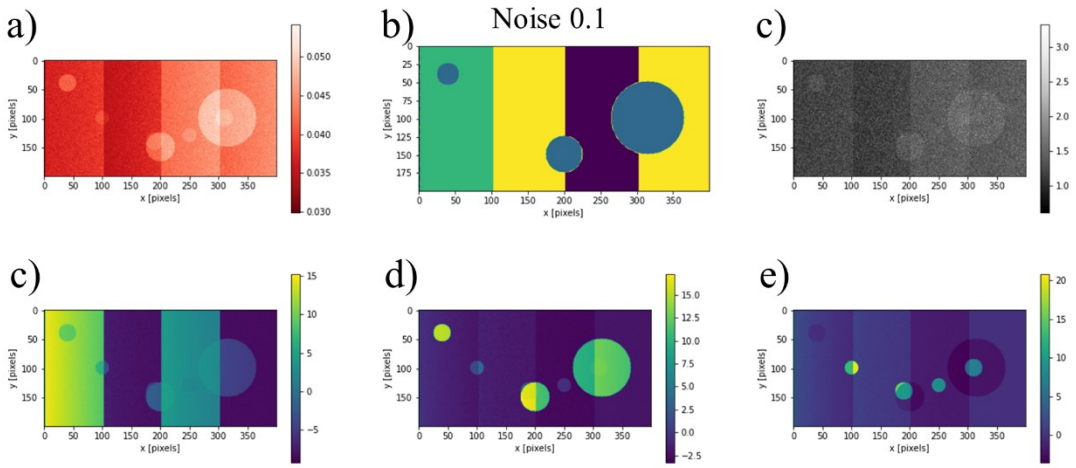

Figure S6: analysis performed with noise up to 1/10 of the maximum signal

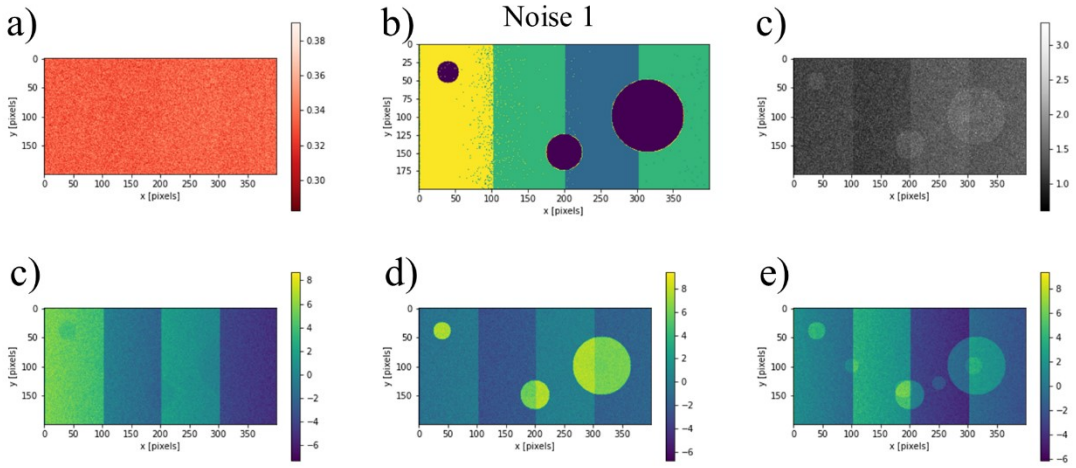

Figure S7: analysis performed with noise up to the maximum signal

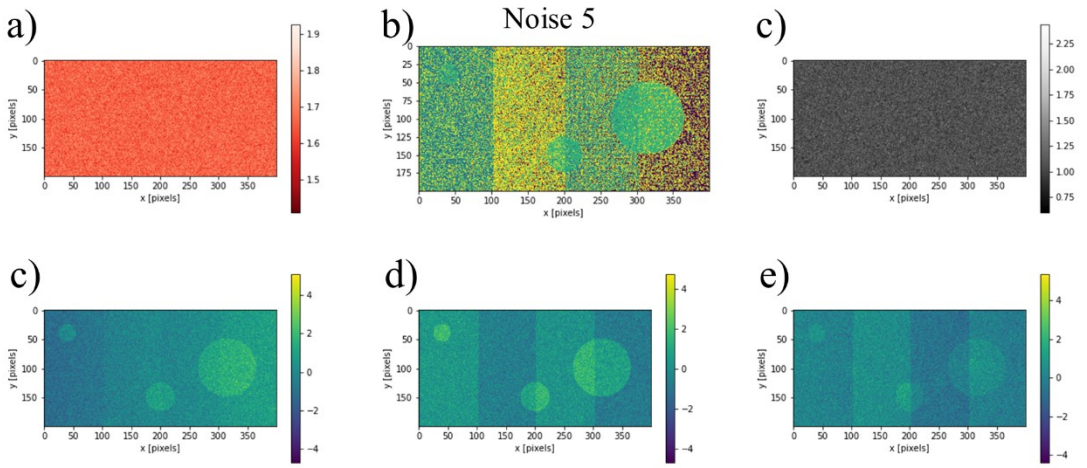

Figure S8: analysis performed with noise up to 5 times the maximum signal

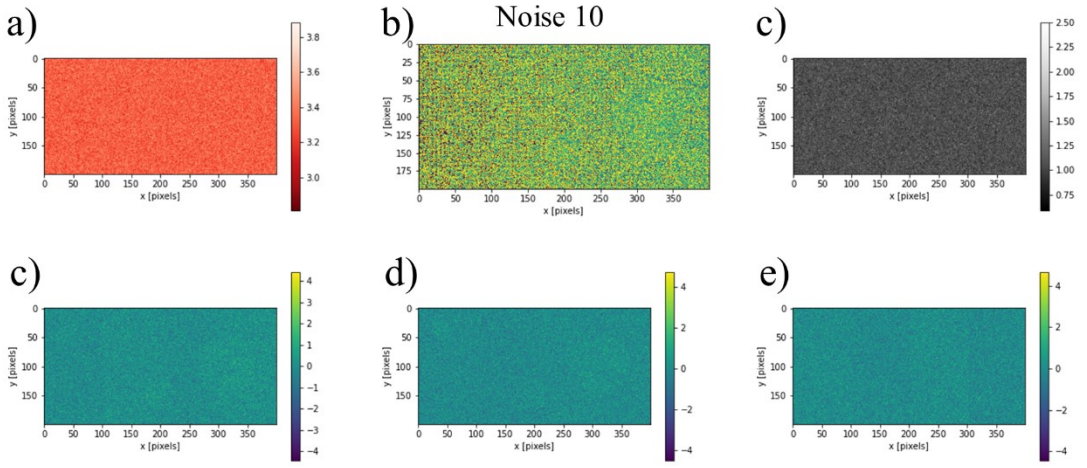

Figure S9: analysis performed with noise up to 10 times the maximum signal

The analytical software is able to identify the four Raman signals up to a maximum noise level equal to the maximum signal, while only one of the signals (Signal 4) is lost when the maximum noise intensity is increased to 5 times the maximum intensity of the Raman signal, and no signal is visible for maximum noise intensities 10 times the maximum Raman signal.

Considering that, in Raman imaging, the spectral quality is usually relatively low due to the necessity to acquire large amounts of data, the noise limit plays a crucial role.

5. Detection limit

To estimate the minimum detection limit in absence of noise, the intensity of Signal 1, Signal 2 and Signal 3 were set to 1, 0 and 1 respectively, while the intensity of Signal 4 was progressively reduced from 1 to 0.01. The results in terms of a) signal intensity b) clustering c) signal over noise ratio, d) PC1 scores e) PC2 scores and f) PC3 scores are presented in Figures S10, S11, S12, S13 and S14.

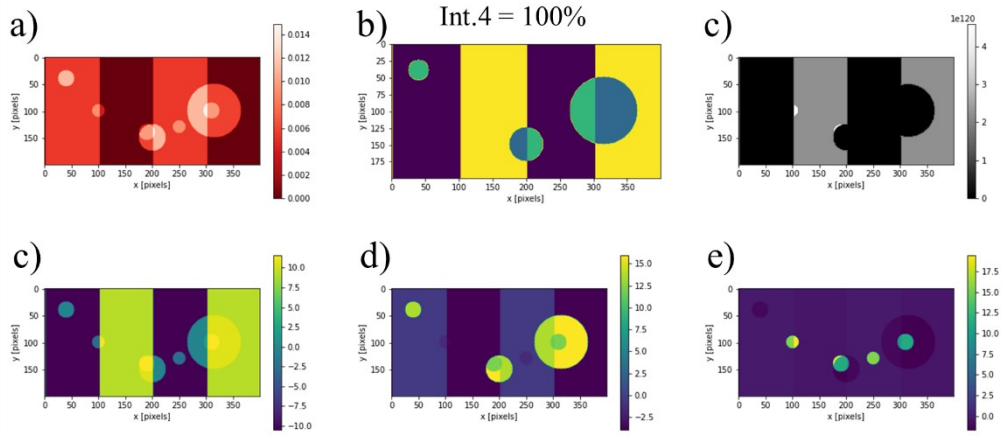

Figure S10: Results for a relative intensity of Signal 4 of 1

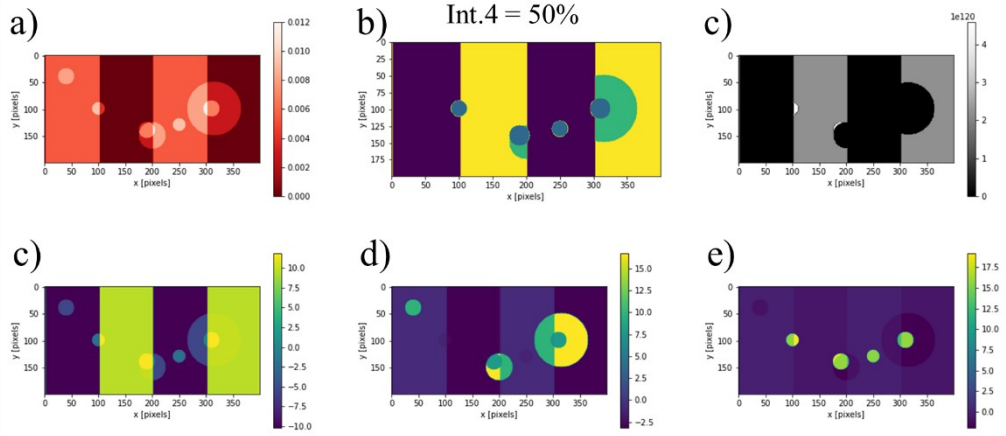

Figure S11: Results for a relative intensity of Signal 4 of 0.5

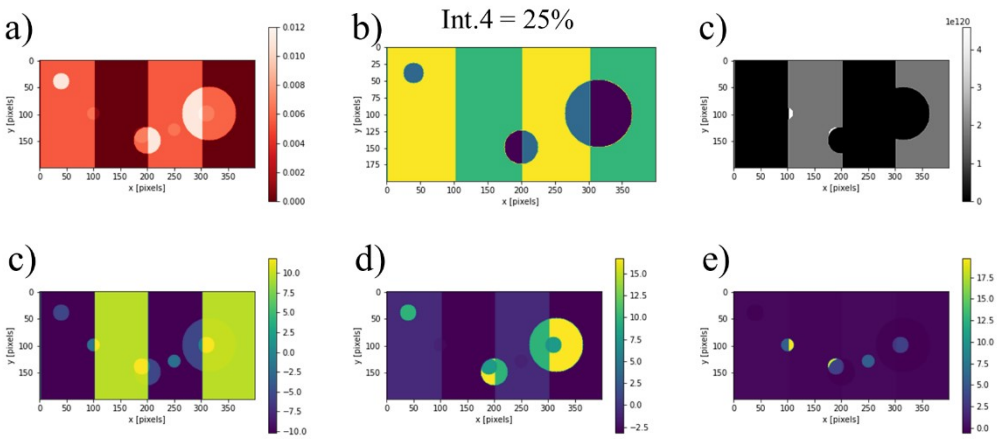

Figure S12: Results for a relative intensity of Signal 4 of 0.25

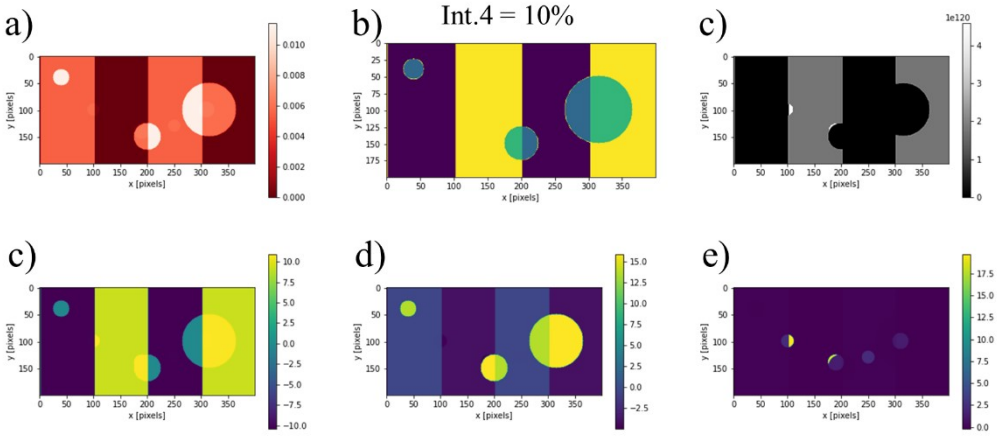

Figure S13: Results for a relative intensity of Signal 4 of 0.1

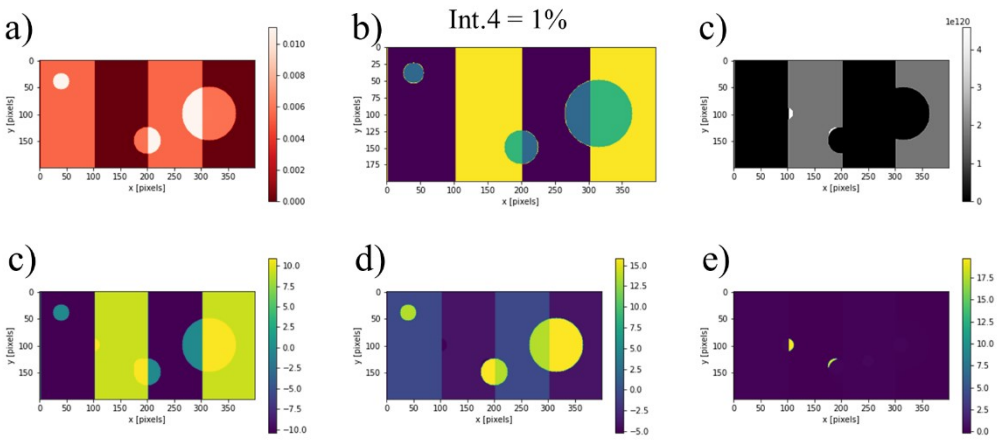

Figure S14: Results for a relative intensity of Signal 4 of 0.01

At a relative intensity of 0.01, the method is unable to correctly discriminate the areas where the Signal 4 is located. It should be noted that the clustering method (subfigures b)) already fails at 0.25.

The presence of noise has also an effect on the detection limit. The limit is unchanged for a noise intensity 0.1 (Fig. S15), but it decreases to 0.25 when the max noise is equal to the max signal (Fig. S16).

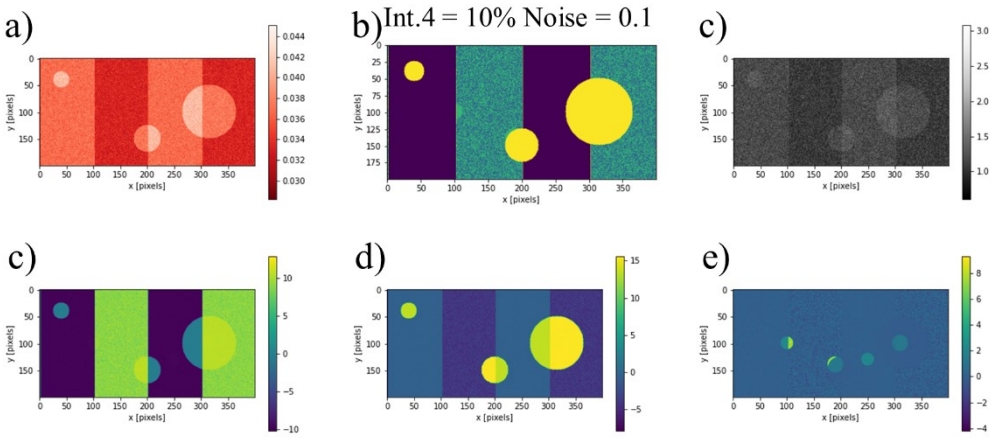

Figure S15: effect of noise on the detection limit, for Signal 4 intensity of 0.1 and noise 0.1

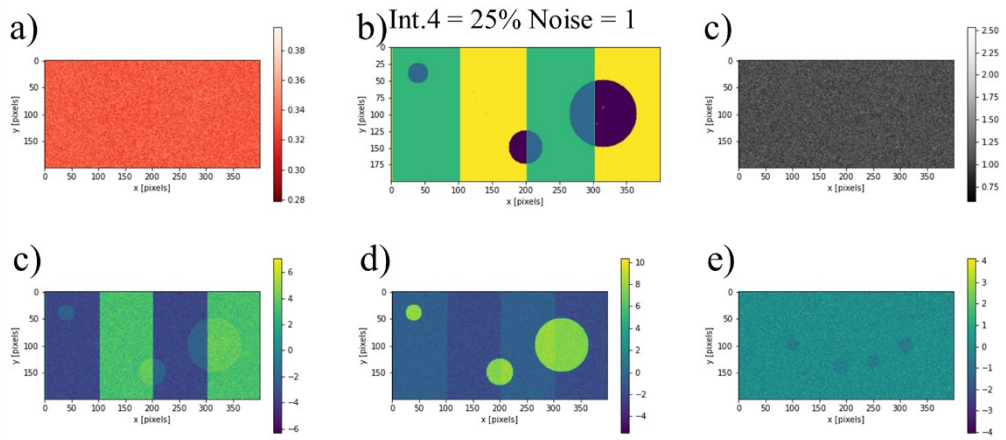

Figure S16: effect of noise on the detection limit, for Signal 4 intensity of 0.25 and noise 1

6. Effect of background

The effect of the background has been evaluated using linear or parabolic curves added to the signals. When the same background is added to all spectra, its effects are negligible as even for background intensities 1,000 times stronger than the Raman signals, the PC maps are not unaffected.

In Raman spectroscopy, the intensity of the background signal varies from point to point. To render this variability in the dummy Raman dataset, we add a random multiplier from 0 to 1 to each background curve before adding it to the signals. Even in absence of noise, the variability of the background intensity is enough to reduce the efficiency of the analysis, and for maximum background intensities 20 times more intense than the weakest signal (Figure S17), it becomes impossible to detect Signal 4.

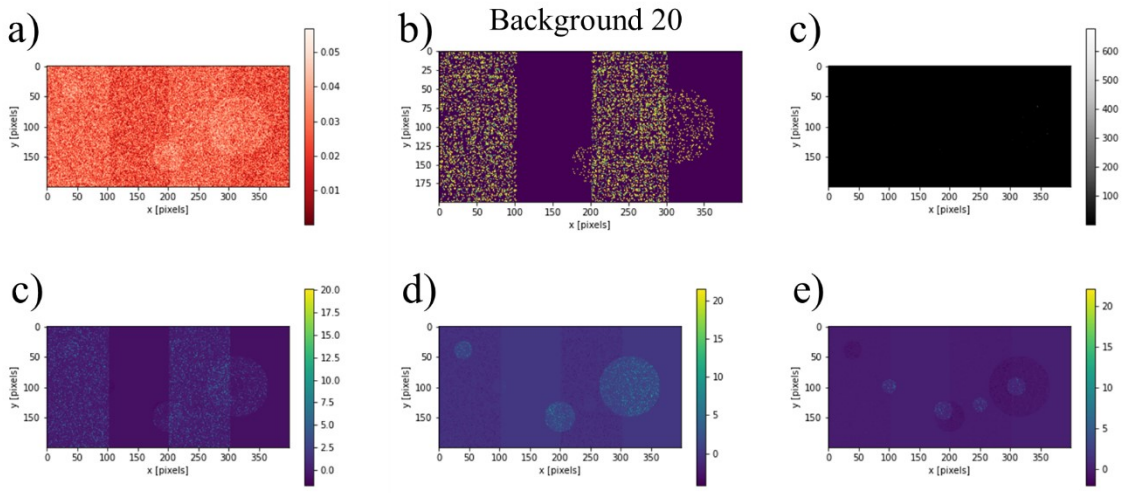

Figure S17: results of the analysis for a background signal 20 times stronger than the Raman signals

This value is further reduced when noise is added to the equations. For a noise value of 0.01 times the signal intensity, the maximum acceptable background intensity multiplier goes from 20 to 10 (Figure S18). For a noise value of 0.1, it further reduces to 6 (Figure S19). For a noise value of 1, it reaches 1.5 (Figure S20).

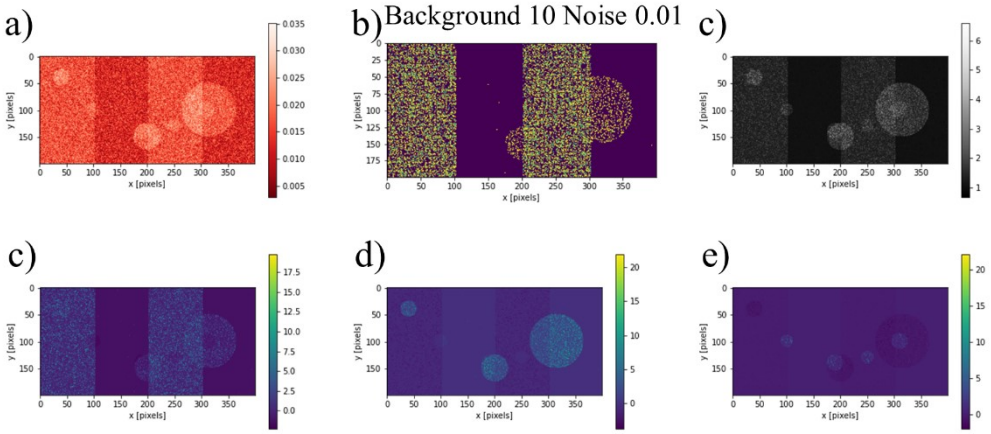

Figure S18: results of the analysis for a background signal 10 times stronger than the Raman signals with a noise relative intensity of 0.01

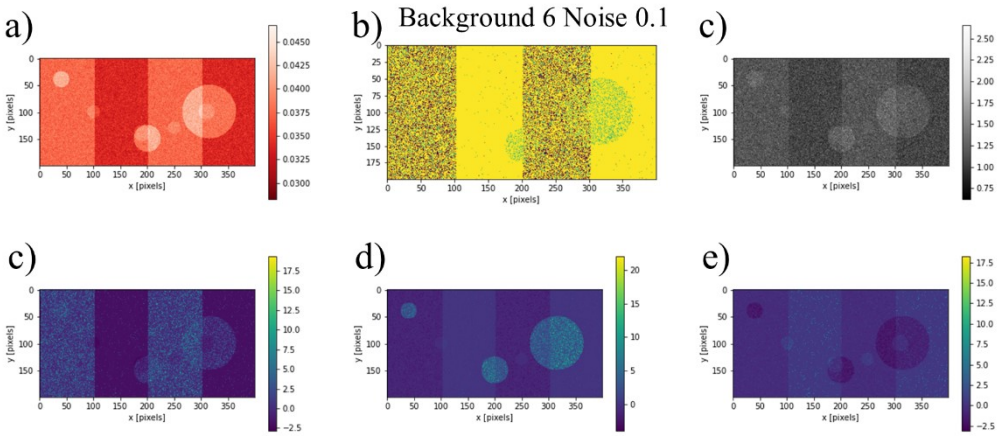

Figure S19: results of the analysis for a background signal 6 times stronger than the Raman signals with a noise relative intensity of 0.1

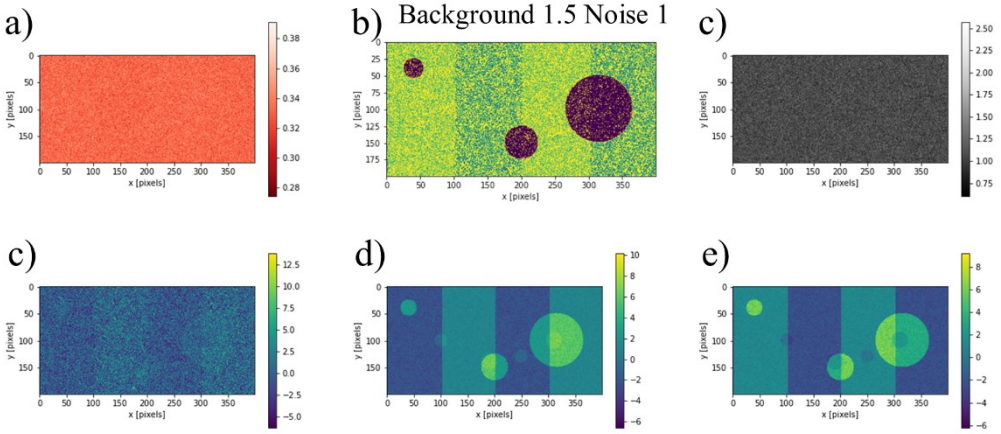

Figure S20: results of the analysis for a background signal 1.5 times stronger than the Raman signals with a noise relative intensity of 1

In real applications, the shape of the background can change along with its relative intensity. This is true in particular for biological samples, so it will be discussed in detail in a follow-up paper focused on Raman spectroscopy for biology.

7. Raw average spectra

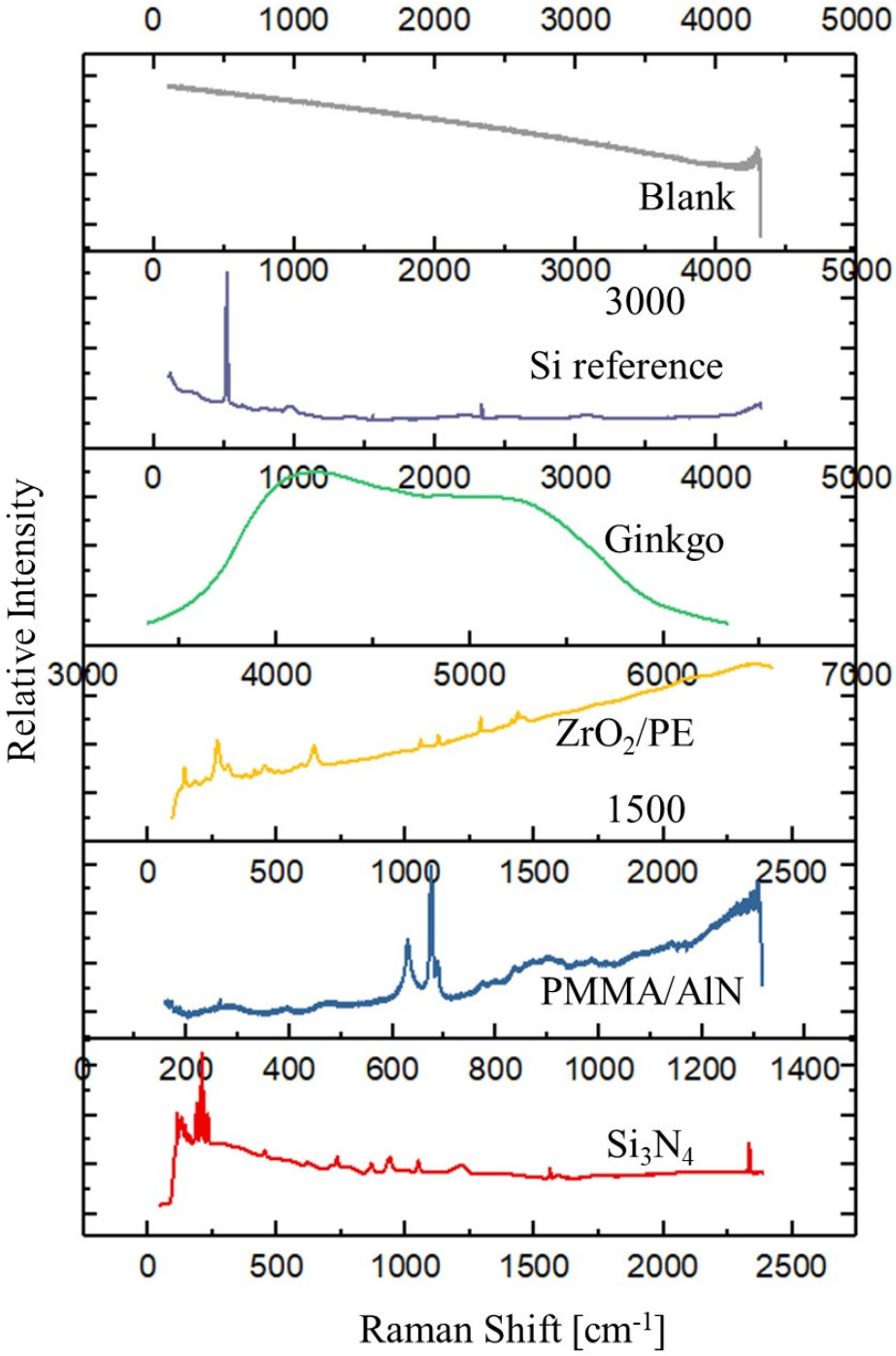

Figure S21: raw average spectra for the six specimen analyzed in the main manuscript

Figure S21 shows the 6 raw average spectra that were utilized during the analytical processes, without any additional pre-treatment such as baseline removal or smoothing.

8. Optimizing the number of clusters

Figure 7 shows a comparison between the *AlN/PMMA* sample and the *ZrO_2/PE* sample for a number of clusters between 1 and 6.

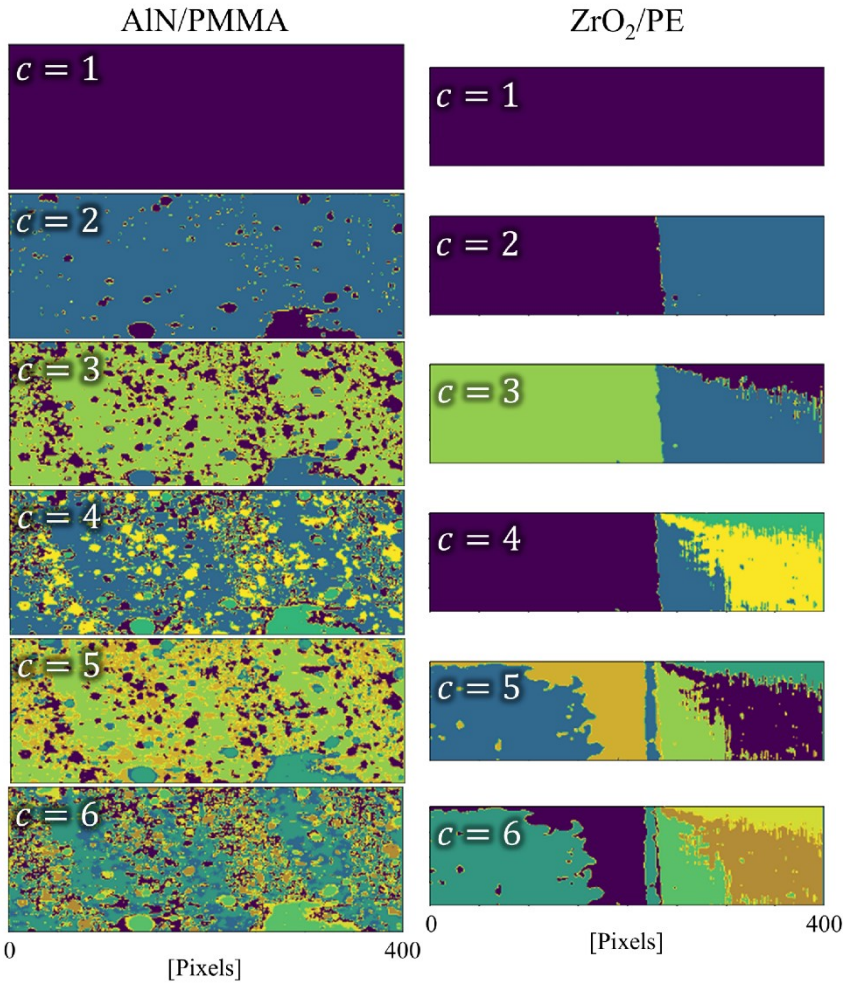

Figure S22: Raman imaging cluster maps for the sample AlN/PMMA (left) and ZrO_2/PE (right) as a function of number of clusters, between 1 and 6.

For samples that possess a homogeneous composition, like the two sides of the *ZrO_2/PE* interface in Figure S22, increasing the number of clusters gives topographical information comparable to contour lines in geographical maps. For samples that also have chemical differences, such as for example because of the presence of secondary phases, the topographical information is still present, but mixed with the chemical mapping to the point that it can be difficult to distinguish between the two. This concept is probably better shown in the *AlN/PMMA* map with c=6. While increasing the number of clusters on the *AlN/PMMA* sample gives better insights about the structure of the sample, to the point that the $c = 6$ map is very similar to the microscope image of Figure 2(e), adding more than 2 clusters

in the *ZrO_2/PE* maps creates sub-regions that only differ because of the alignment of the surfaces, giving no real spectroscopic clue for further analysis.

9. PC2 *vs* PC3 and PC1 *vs* PC3 scores

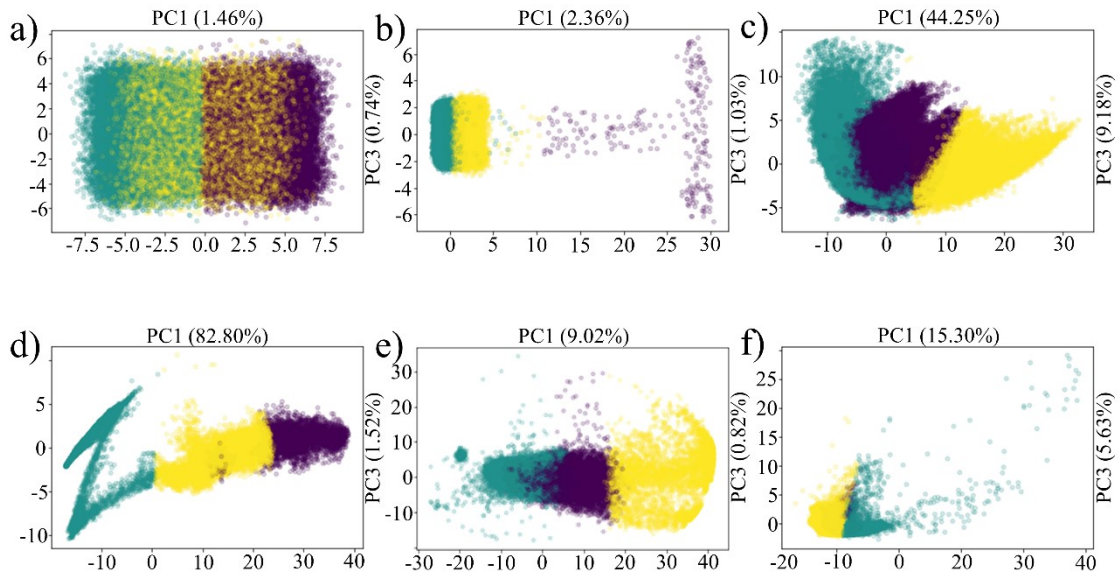

Figure S23: PC2 and PC3 scores for the six different samples: (a) blank, (b), Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4

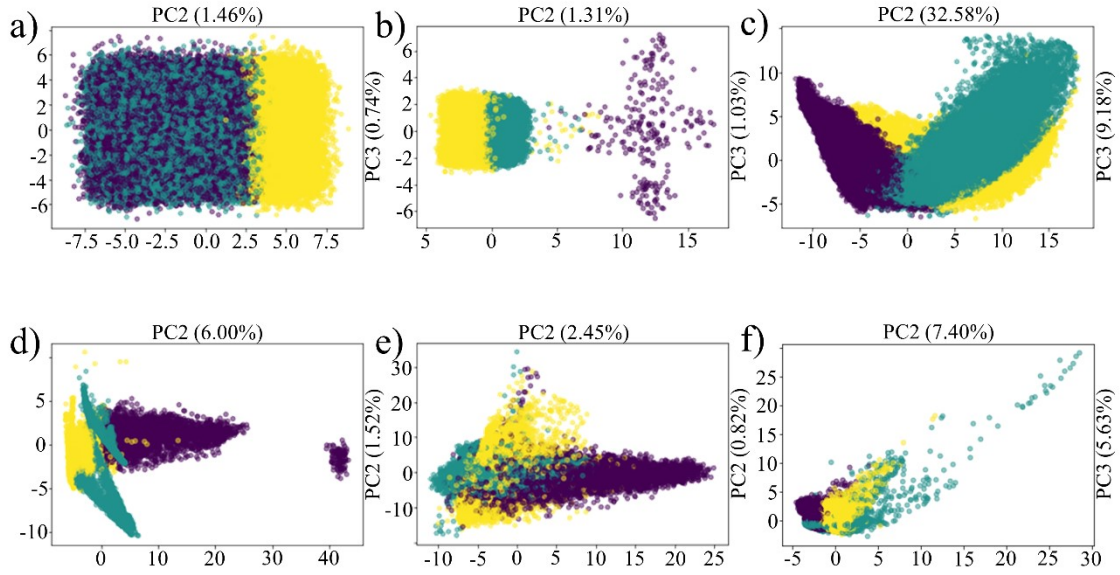

Figure S24: PC1 and PC3 scores for the six different samples: (a) blank, (b), Si reference, (c) ginkgo, (d) ZrO_2/PE, (e) AlN/PMMA, (f) Si_3N_4

Figure S23 and Figure S24 show the scatter plots related to the scores of PC2 *vs* PC3 and PC1 *vs* PC3. In four cases, the Blank sample, the Si reference, the ZrO_2/PE interface and the AlN/PMMA composite, the scores of PC3 resulted to be lower than 2%, making the contribution of the third principal component virtually irrelevant. This is further confirmed by the strong overlapping between different clusters in the relative scatter plots. Both the Ginkgo leaf and the Si_3N_4 block see a substantial contribution of the third principal component, and the reason behind this will be further discussed in the main manuscript.