Check for updates

OPEN

New methodology to process shifted excitation Raman difference spectroscopy data: a case study of pollen classification

F. Korinth[1], A. S. Mondol[1], C. Stiebing[1], I. W. Schie[1,2], C. Krafft[1] ✉ & J. Popp[1,3]

Shifted excitation Raman difference spectroscopy (SERDS) is a background correction method for Raman spectroscopy. Here, the difference spectra were directly used as input for SERDS-based classification after an optimization procedure to correct for photobleaching of the autofluorescence. Further processing included a principal component analysis to compensate for the reduced signal to noise ratio of the difference spectra and subsequent classification by linear discriminant analysis. As a case study 6,028 Raman spectra of single pollen originating from plants of eight different genera and four different growth habits were automatically recorded at excitation wavelengths 784 and 786 nm using a high-throughput screening Raman system. Different pollen were distinguished according to their growth habit, i.e. tree versus non-tree with an accuracy of 95.9%. Furthermore, all pollen were separated according to their genus, providing also insight into similarities based on their families. Classification results were compared using spectra reconstructed from the differences and raw spectra after state-of-art baseline correction as input. Similar sensitivities, specificities, accuracies and precisions were found for all spectra with moderately background. Advantages of SERDS are expected in scenarios where Raman spectra are affected by variations due to detector etaloning, ambient light, and high background.

Raman spectroscopy is a vibrational spectroscopy technique that is used for the assessment of the chemical composition of samples. Even complex biological samples can be analyzed in a non-destructive and label-free manner and classified using their specific molecular fingerprints assessed by this method[1–4]. However, intense and strongly varying backgrounds, e.g. due to autofluorescence (with or without photobleaching), detector etaloning effects and ambient light, are an often occurring challenge in Raman spectroscopy. If the Raman intensity is too low relative to the background intensity, Raman bands are hard to discern or are masked completely. Although it cannot be excluded that an autofluorescence background contains useful information, background correction procedures are state-of-art in Raman spectroscopy of biological material. Autofluoerescence contributions in Raman spectra seem to be sensitive, but specificity might be problematic due to bleaching and quenching effects, which are prone to variations and lack proper reproducibility.

Therefore, different approaches were suggested to tackle the challenge of intense and varying backgrounds. One option for autofluorescence is the destruction of the fluorophores by photobleaching[5–7], which is rather time-consuming and might cause side effects, such as sample contamination with chemiphotobleaching agents or thermal stress due to extended laser exposure of the sample. A review divided other techniques roughly into two groups: computational and instrumental background correction methods[8]. Examples of typically used computational background correction algorithms are extended multiplicative signal correction[9,10] (EMSC), multiplicative signal correction[11] (MSC), rubberband[12,13], sensitive nonlinear iterative peak[14] (SNIP), and polynomial fittings[15]. These approaches often require high computational effort and need experienced personal for the data analysis. Cordero et al. corrected a high fluorescence background in Raman spectra of bladder biopsies using EMSC[16]. For in vivo Raman spectra of colorectal tissue Bergholt et al. corrected autofluorescence background by a high-order

[1]Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany. [2]Department of Medical Engineering and Biotechnology, University of Applied Sciences, Carl-Zeiss-Promenade 2, 07745 Jena, Germany. [3]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany. ✉email: Christoph.krafft@leibniz-ipht.de

polynomial fitting[17]. The in vivo acquired Raman spectra of brain cancer by Desroches et al. were also background corrected using a polynomial[18]. Galli et al. found a high fluorescence background in Raman spectra of brain biopsies, where 88.4–96.5% of the collected intensities were attributed to fluorescence. For separating the background-free Raman signal and the fluorescence profile a baseline estimation toolkit was used. The authors concluded that the classification was best, when both information were used[19].

Instrumental background correction methods for fluorescence rejection are time-gating approaches, where the fast Raman scattering is detected before the slower fluorescence emission, and phase or wavelength modulated techniques, where the Raman scattering changes according to the wavelength or phase modulation whereas the fluorescence emission does not[8]. Another promising method is shifted excitation Raman difference spectroscopy (SERDS)[20].

SERDS belongs to the instrumental background correction methods, which uses two slightly shifted excitation wavelengths to acquire two Raman spectra consecutively at the same lateral position. The shift in excitation wavelength is chosen to be only a few nanometers, leading to two slightly shifted Raman spectra with the same fluorescence background profile, since the same fluorophores are excited. After subtraction of the shifted Raman spectra from each other, the resulting difference spectrum is ideally free of background contributions and only contains Raman information. Furthermore, other constant spectral contributions such as ambient light or the system transfer function (e.g. detector etaloning effects) can be suppressed[21]. Proof of principle studies demonstrated SERDS using several combinations of solvents and dyes as model analytes, especially for the introduction of new lasers with two or more excitation wavelengths[20,22–25]. Sowoidnich and Kronfeldt analyzed different laser wavelengths for SERDS experiments on parts of beef and pork tissues like fat, connective tissue, bone and meat[26]. Noack et al. conducted SERDS measurements to measure algae cultivation samples and monitor sulfated exopolysaccharides (EPS) concentrations in the reactors. For this, 10 raw spectra were averaged, smoothed and baseline corrected before the subtraction. A principle component analysis (PCA) and different regression models were then applied to the smoothed difference spectra to determine the EPS concentration, which worked poorly for the partial least squares regression (PLSR) model, but very well for the support vector regression (SVR) model[27]. Martins et al. studied molar teeth ex vivo and human skin in vivo using SERDS with an excitation wavelength of 830 nm/830.5 nm and regular Raman spectroscopy at 1,064 nm as a control. For data analysis the difference spectra were integrated to reconstruct the Raman spectrum[28]. Gebrekidan et al. measured difference spectra of pig tissue (bone, fat, gland and mucosal). After a sophisticated data processing including normalization, first baseline correction, reconstruction and second baseline correction to receive fluorescence-free pure Raman spectra, a classification by PCA was performed[29]. By measuring a plate of clear polystyrene, Maiwald et al. showed that SERDS was able to filter out ambient light passing through polystyrene. They also conducted SERDS in an orchard measuring the wax on the skin of an apple and the chlorophyll in a leaf using a handheld device with a high numerical aperture[30,31]. Schmälzlin et al. obtained SERDS images from different samples, e.g. cross-section of a pig ear, skin and a dissolving brown sugar cube using a custom-built multi-focus probe head and an integral field spectrograph. This system was able to simultaneously detect 400 spectra delivered by the probe head of 20×20 pixels[32].

Since photobleaching and intensity variations due to e.g. laser power and filter characteristics often result in varying background intensities, most difference spectra are not completely background-free. This makes additional background correction steps necessary. Also reconstruction steps are usually implemented to transform the difficult to interpret difference spectra into accustomed Raman spectra. There are several reconstruction approaches, such as deconvolution, linear data manipulation, integration, kernel function, or non-negative least squares fitting[33–34]. These reconstruction methods always harbor the risk of introducing artefacts into the reconstructed Raman spectrum, due to the correlation between the fixed wavelength shift and varying Raman band widths[35].

As a case study pollen samples of eight different plant genera were investigated. Pollen are a valuable case study, since their Raman spectra experience intensity differences in the fluorescence backgrounds (see supplementary information in Ref.[36]). In palynology, pollen are taxonomically evaluated under a microscope considering their morphology. This is time-consuming and requires a highly trained expert to differentiate several hundreds of different pollen. There are several ideas for automatization and technical improvement of this gold standard[37–39]. Other spectroscopic approaches like infrared absorption[40–42], laser-induced breakdown[43] or Raman scattering have been applied[35,43–55]. Raman spectroscopy was implemented to build a spectral database of pollen including a chemometrical classification by their growth habit[36].

The approach in this work to differentiate several pollen genera uses difference spectra, that were obtained by novel processing of SERDS data, and lends itself as a case study to classify biological samples. The new streamlined method to handle and classify SERDS data of biological samples is based on their single difference spectra without a reconstruction step to retrieve the familiar profile of Raman spectra or baseline correction procedures. Furthermore, spectra were reconstructed from the differences and Raman spectra were processed by state-of-art baseline correction. For classification using difference spectra, reconstructed spectra and baseline corrected spectra as input, a PCA followed by a linear discriminant analysis (PCA-LDA) was chosen. The classification results were compared with respect to sensitivities, specificities, accuracies and precisions.

Results and discussion

Data acquisition and processing.
Representative raw spectra of a single birch and hazel pollen grain at three consecutively measured excitation wavelengths of $\lambda_1 = 784$ nm (130 mW), $\lambda_2 = 785$ nm (180 mW) and $\lambda_3 = 786$ nm (200 mW) are shown in Fig. 1a. The series of consecutive measurements started with 784 nm followed by 785 nm and 786 nm. The spectra show similarities in e.g. amide bands (1,310 and 1,650 cm^{-1}) and the sporopollenin bands (1,007, 1,454 and 1614 cm^{-1}) but vary in intensity. When separately evaluating the two sets
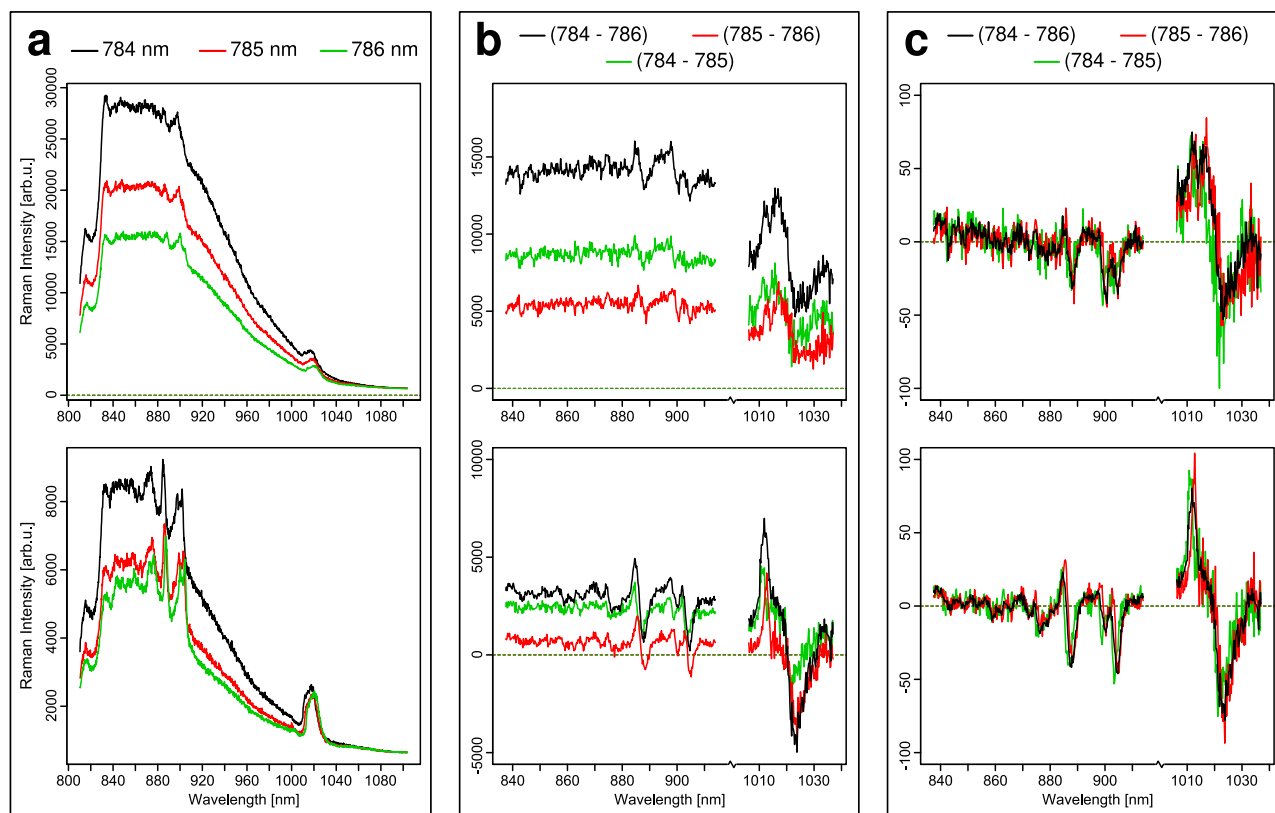
Figure 1. Overview of the data processing steps. top row, spectra of one birch pollen grain; bottom row, spectra of one hazel pollen grain; (**a**) raw spectra at the three different excitation wavelengths; (**b**) raw difference spectra; (**c**) normalized and optimized difference spectra; the color codes of excitation wavelengths are indicated on top.

of Raman spectra, the Raman signals shift by approx. 16 cm^{-1} per 1 nm wavelength shift. Although the laser intensities increased from 130 to 180 mW and 200 mW for 784 nm, 785 nm, and 786 nm, the signal intensities decrease in the same order. This is due to photobleaching of the autofluorescence background. Therefore, the decrease of spectral background during the onset of the measurements is stronger than the increase of the Raman bands due to elevated laser intensities. Figure 1b shows the three resulting difference spectra. For the red and green difference spectra a 1 nm shift (green: 784–785 nm; red: 785–786 nm) was realized, whereas the black difference spectrum resulted from a 2 nm shift. A significant, variable offset remained between the difference spectra, clearly originating from the varying fluorescence profile, which is maximum for the 2 nm shift. Figure 1c shows the normalized and optimized difference spectra (see data preprocessing in the methods section) of aforementioned difference spectra. The background is successfully corrected. Closer inspection reveals that a 1 nm shift results in difference spectra with higher noise than the 2 nm shift due to smaller amplitudes in the raw differences (see Fig. 1b). The shift in excitation wavelength should be near the full width at half maximum (FWHM) of the Raman band for the proper interpretation and reconstruction of a Raman spectrum. But since Raman bands have different FWHM as can be seen in Fig. 1a, there is not one wavelength shift that fits all FWHM of the Raman bands. Furthermore, technical parameters such as the transmission range of the laser line filter limit the wavelength shift to 2 nm.

Therefore, further data analysis was performed on the wavelength pair 784–786 nm. A trend is evident that negative difference bands are more intense than positive difference bands, which is a consequence of optimization step and the higher laser intensity at 786 nm. In the optimization step the subtrahend is multiplied by a factor to compensate for differences in background intensity. Since the subtracted spectra with the excitation wavelength of 786 nm always have a lower background due to photobleaching, all spectra are multiplied by an optimization factor larger than 1 resulting also in higher peak intensity and therefore more intense negative difference peaks.

Figure 2 gives an overview of the mean spectra (dark) and their respective standard deviation (shaded). Figure 2a shows the four tree pollen difference spectra and Fig. 2b the four non-tree pollen difference spectra for the 2 nm shift, that constitute the basis for the following classification. Figure 2c, d show the reconstructed Raman spectra after baseline correction, and Fig. 2e, f the raw Raman spectra ($\lambda_{ex} = 784$ nm) after baseline correction for comparison. Spectral differences in the tree pollen occur between 1,600 and 1,750 cm^{-1}, and in the high wavenumber region for larch, most likely due to higher lipid contributions of the conifer typical essential oils. Differences in non-tree pollen are also observed between 1,600 and 1,750 cm^{-1} and additionally in the low wavenumber region below 1,200 cm^{-1}, which can be explained by their different families and growth habits (i.e. herb, grass, and shrub). One exception is rumex and cyclamen, which show a distinct difference in band structure although they stem from the same growth habit. In moor grass, the band near 1,600 cm^{-1} is weak. The
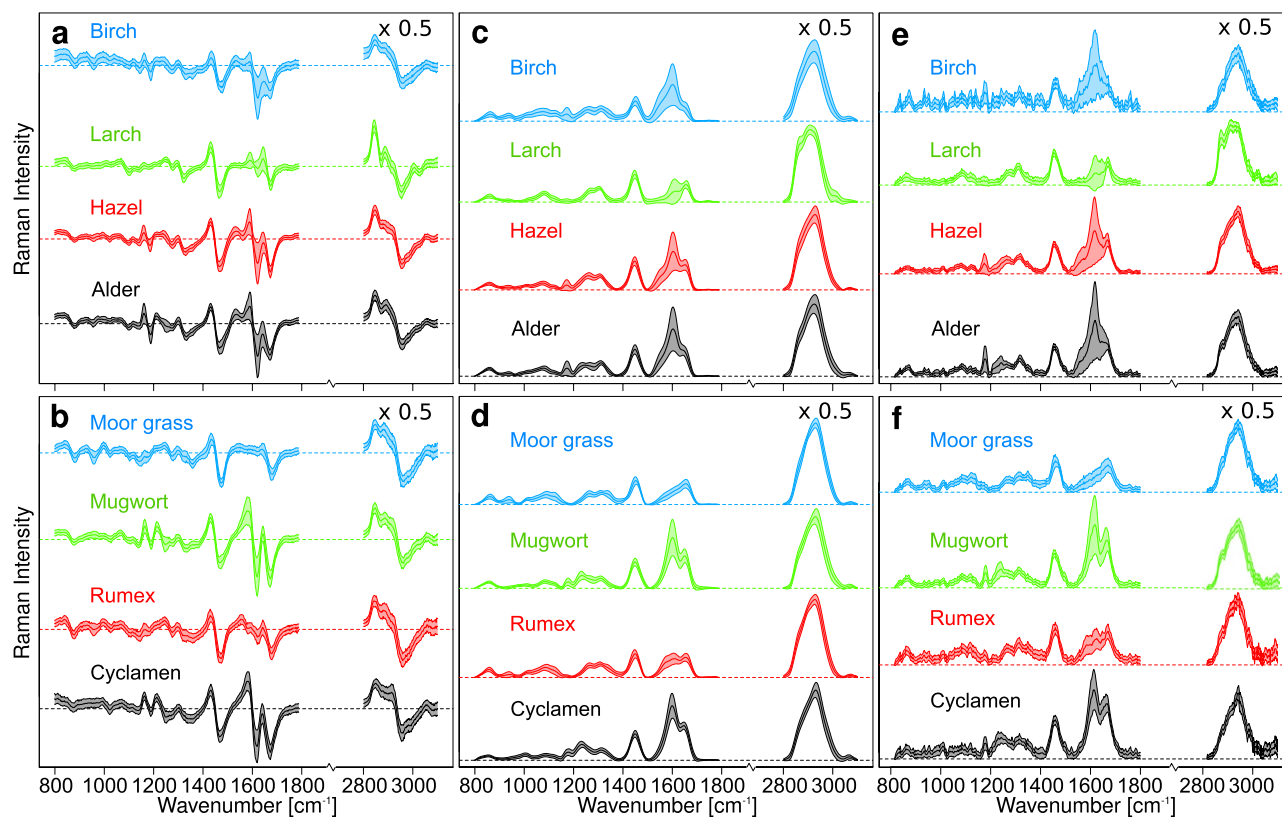
Figure 2. Mean and standard deviation of difference spectra (784–786 nm) after normalization and optimization for tree pollen (**a**) and non-tree pollen (**b**), spectra reconstructed from differences after SNIP baseline correction of tree pollen (**c**) and non-tree pollen (**d**) and Raman spectra ($\lambda_{ex} = 784$ nm) after SNIP baseline correction for tree pollen (**e**) and non-tree pollen (**f**). High wavenumber regions were multiplied by 0.5.

standard variations of the $1,600$ cm^{-1} band is high for all pollen. In particular, some Raman spectra of birch, larch, hazel, alder and rumex pollen also have weak intensities near $1,600$ cm^{-1}. The reconstruction results in spectra that properly mimic the raw Raman spectra after both data were baseline corrected using the SNIP algorithm. The main difference between reconstructed and raw spectra is that the reconstruction algorithm reduces the spectral resolution. It is important to note that the subtraction algorithm does not alter the spectral resolution of the difference spectra.

Classification of tree vs non-tree pollen data. In the first classification step the whole data set was divided into a training data set and a test data set. A PCA was performed on the whole training data set of all pollen samples to separate major variations in the lower principal components (PCs) from noise in the higher PCs. Supplementary Fig. S1 shows the loadings of the first 15 PCs and the loading of the LDA model. Only the first 10 PCs accounting for ca. 68.6% of the variance were used for training the LDA model (for the explained variance curve see Supplementary Fig. S2). Intense variations in the loadings of the first 10 PCs can easily be seen. Almost no information is anymore provided in PC 14 and higher, whereas the high wavenumber region is dominated by noise starting from PC 11 due the lower quantum efficiency of the detector and hence lower signal intensities. A band assignment is not straightforward in the case of PC loadings based on difference spectra, since their variations are related to positive and negative Raman difference bands. This is also the case for the LD loading presented in Supplementary Fig. S1d. In the high wavenumber region the CH$_3$ stretching (symmetric and asymmetric) bands, the CH$_2$ stretching (symmetric and asymmetric) bands and the CH stretching bands overlap to a very broad convoluted band structure. Due to the 2 nm shift in excitation wavelength a lot of signal intensity is lost in the difference spectrum as the bands are shifted into each other. This, combined with the lower quantum efficiency of the detector, leads to a LD1 spectrum with more noise in the high wavenumber region.

The LD scores of the separation between tree pollen and non-tree pollen are shown in the box and whiskers plots in Fig. 3. In the prediction of the test data, all scores with a negative LD1 score belong to the non-tree class, all scores with a positive LD1 score belong to the tree class. Some misclassifications can be seen for the real classes. However, the medians and 0.25/0.75 quantiles are well separated and away from the class boundary at 0.

The confusion matrix is provided in Table 1. High sensitivity, specificity, accuracy and precision of over 95% are achieved with the constructed PCA-LDA model on average for all classifiers. The two classes can be very well separated using the normalized and optimized difference spectra and the constructed PCA-LDA model. Of the 133 misclassified tree pollen 22% were alder, 28% hazel and 50% birch pollen. Larch pollen contain a lot
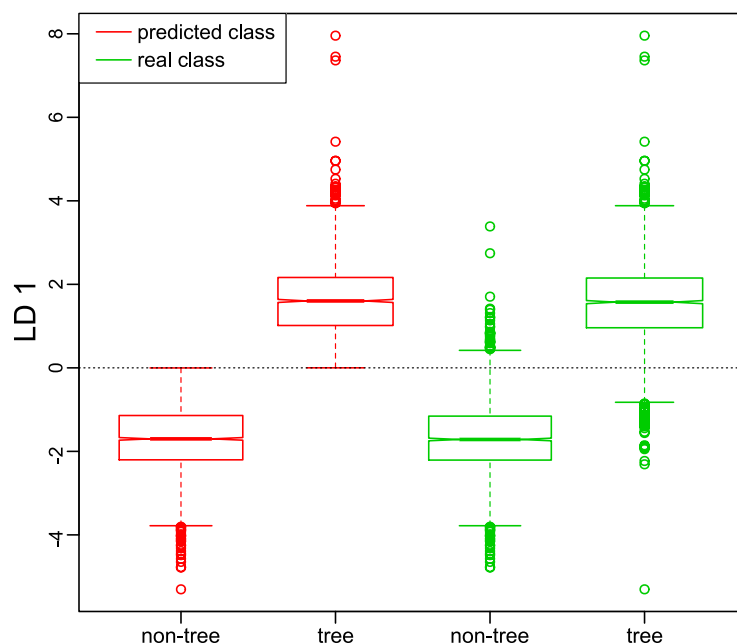
Figure 3. Box and whiskers LDA score plot for classification of tree vs. non-tree. Red, predicted classes of the scores of the test data set; green, real classes of the scores of the test data set.

Tree vs. non-tree	Real classes	
	Non-tree	Tree
Non-tree, predicted	2,700	133
Tree, predicted	86	2,371
Sensitivity	96.9	94.7
Specificity	94.7	96.9
Accuracy	95.9	95.9
Precision	95.3	96.5

Table 1. Confusion matrix of non-tree versus tree classification: real classes vs. predicted classes of the classified test spectra; sensitivity, specificity, accuracy and precision in %.

of lipids due to the conifer-typical essential oils, as also indicated in the difference spectra in Fig. 2, which leads to a clear separation without any misclassification.

Classification of the different pollen genera. To further analyze the feasibility of classification schemes based on SERDS difference spectra, the data set was separated into their simplified growth habits, i.e. tree and non-tree data, to classify each group into their genera. Again, the data sets were split into training and test, and a PCA-LDA model including internal cross validation was implemented for each group. The LD loadings for the discrimination of the different tree pollen types are shown in the Supplementary Fig. S3a and for the different non-tree pollen types in the Supplementary Fig. S3b. As before, the loadings and their interpretation are very complex. Nevertheless, LD3 for the non-tree separation shows a less intense spectrum. Consequently the noise has a much higher impact on the spectrum, which can be seen especially for the high wavenumber region.

For LDA modelling of the different tree pollen types, the first 13 PCs were used explaining a cumulative variance of 78.6% (for the explained variance curve see Supplementary Fig. S4). The LD scores for the different tree pollen types were successfully separated, which is shown in the LD score plot (Fig. 4). LD1 separates larch from the other trees, LD2 separates birch from the other trees and LD3 grossly separates alder from hazel. The 3D score plot shows the separation of the scores of each type.

The classification results of the different tree pollen genera are summarized in Table 2. Especially the rates for the larch pollen samples show the best values: a specificity of 99.8% and an accuracy of 99.7% are the highest values for all pollen sample classifications. This is not surprising since larch is from a different family (Pinaceae) than the other three tree types (Betulaceae). Due to the high lipid content, larch pollen did not sediment very well onto the substrate causing an overall low number of automatically detected larch pollen samples for Raman
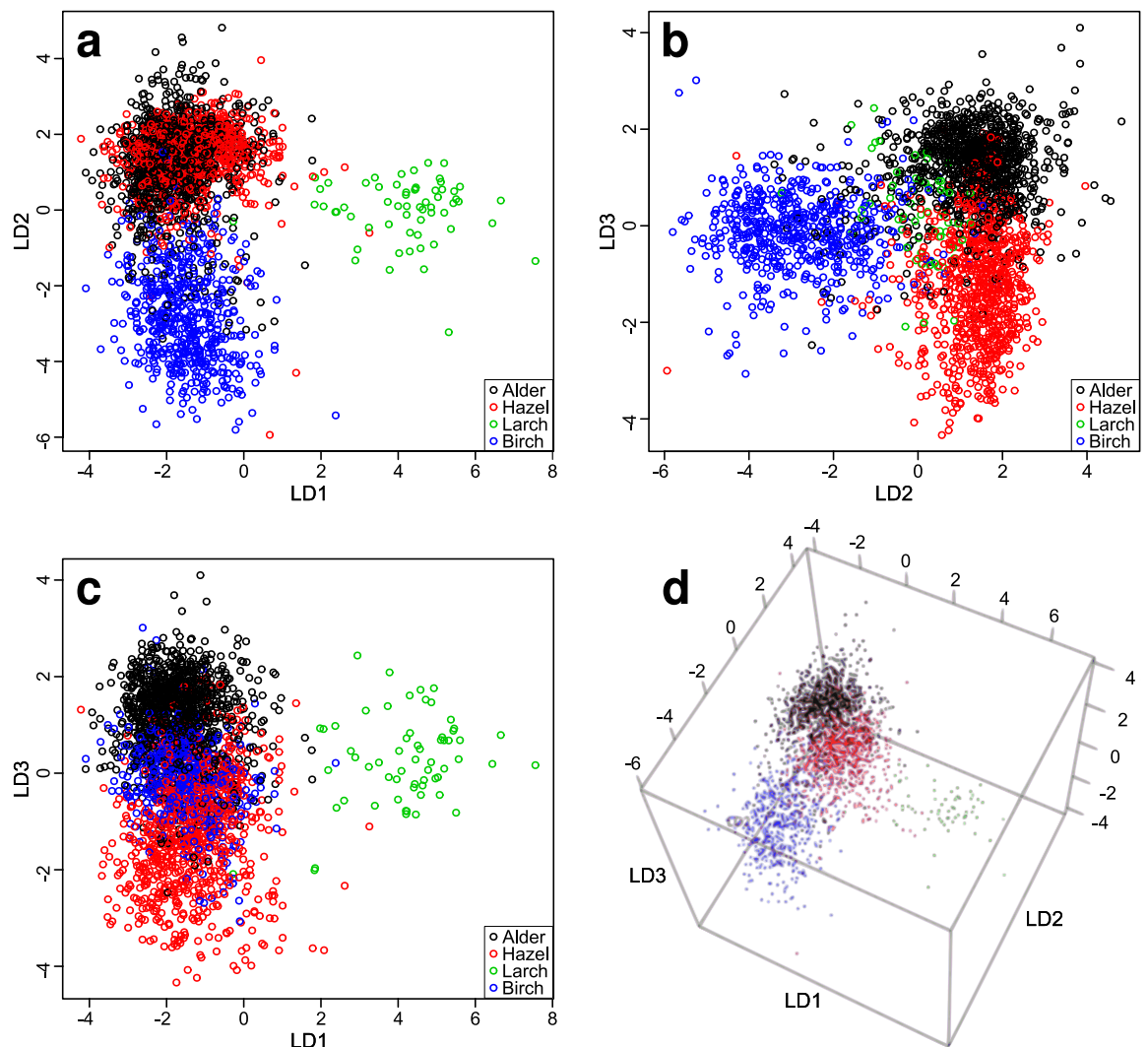
Figure 4. LDA score plots for classification and separation of different tree pollen genera. (**a**) LD1–LD2 plane of the score plot; (**b**) LD2–LD3 plane of the score plot; (**c**) LD1–LD3 plane of the score plot; (**d**) 3D score plot.

	Real classes			
	Alder	Hazel	Larch	Birch
Alder, predicted	947	107	0	19
Hazel, predicted	54	696	2	16
Larch, predicted	3	2	59	0
Birch, predicted	52	14	0	533
Sensitivity	89.7	85.0	96.7	93.8
Specificity	91.3	95.7	99.8	96.6
Accuracy	90.6	92.2	99.7	96.0
Precision	88.3	90.6	92.2	89.0

Table 2. Confusion matrix for classification and separation of tree pollen types. Upper part: real classes vs. predicted classes of the classified test spectra; lower part: sensitivity, specificity, accuracy and precision in %.

spectra acquisition (in total 61 spectra). For the separation of the tree pollen types, approx. 90% of alder pollen, 85% of hazel pollen, 97% of larch pollen and 94% of birch pollen were correctly classified.
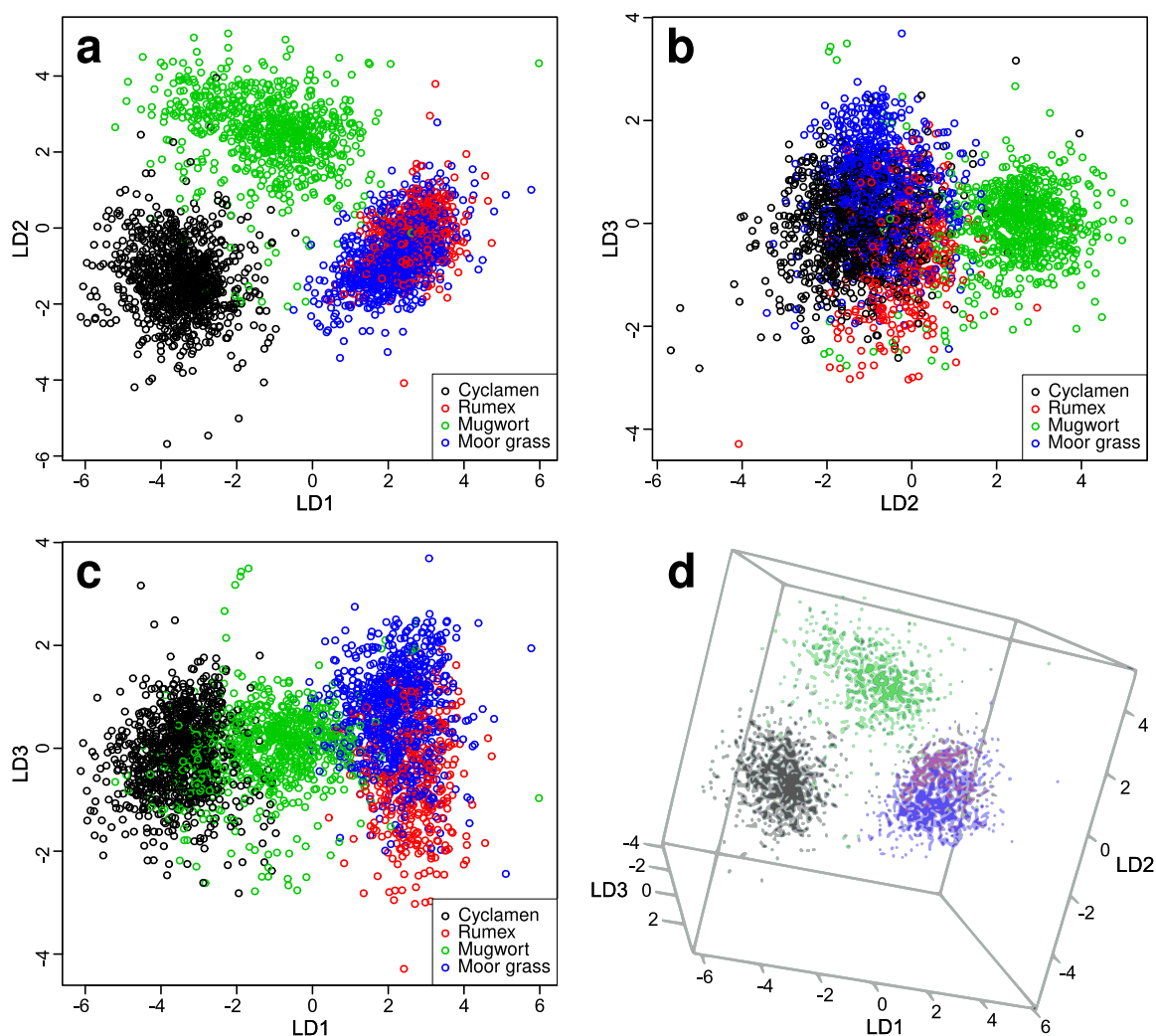
Figure 5. LDA score plot for classification and separation of different non-tree pollen types. (**a**) LD1–LD2 plane of the score plot; (**b**) LD2–LD3 plane of the score plot; (**c**) LD1–LD3 plane of the score plot; (**d**) 3D score plot.

For the discrimination of the non-tree data set 11 PCs were included after an internal cross validation of the training data set, explaining 61.8% of the cumulative variance (the explained variance curve see Supplementary Fig. S5). The PCA-LDA model was then used for the classification of the test data set into the four different non-tree types: mugwort, cyclamen, moor grass and rumex. The score plots of the test data classification are shown in Fig. 5. The cyclamen scores are separated from moor grass by LD1, and mugwort from all other non-tree pollen by LD2. The rumex scores can be best separated by LD3, but still exhibit a significant overlap with moor grass. In the 3D plot the scores of cyclamen and mugwort are very well separated from the other two classes, whereas rumex and moor grass have a larger overlap, thus making it hard to separate the two pollen types from one another. Table 3 shows the results of the classification of the non-tree pollen types and the corresponding classifiers in form of a confusion matrix.

In the separation between the different non-tree pollen, 98% of the cyclamen pollen, 74% of the rumex pollen, 94% of the mugwort pollen and 79% of the moor grass pollen were correctly classified. Rumex and cyclamen are herbs, mugwort a shrub type plant and moor grass a grass. The SERDS spectra of rumex and moor grass pollen are quite similar resulting in the most misclassifications. The classification of grass pollen genera by Raman spectroscopy was challenging which was also found by Mondol et al.[36].

Comparison with classification of reconstructed and raw Raman spectra. Reconstructed and raw Raman spectra after baseline correction as shown in Fig. 2 for tree and non-tree pollen were subjected to the analogous PCA-LDA classification. Confusion matrixes are presented in Supplementary Table S2. To simplify the comparison, average values for sensitivity, specificity, accuracy and precision were calculated and displayed in Supplementary Table S1. The classification rates of tree versus non-tree pollen agreed well for difference spectra and spectra at 784 and 786 nm excitation, but are 2–3% lower for the reconstructed spectra. The classification rates of tree types varied between 88 and 98%, and the reconstructed spectra tended to give lower values than

	Real classes			
	Cyclamen	Rumex	Mugwort	Moor grass
Cyclamen, predicted	897	0	18	0
Rumex, predicted	0	267	16	168
Mugwort, predicted	16	0	676	0
Moor grass, predicted	0	96	10	622
Sensitivity	98.2	73.6	93.9	78.7
Specificity	99.0	92.4	99.2	94.7
Accuracy	98.8	89.9	97.8	90.2
Precision	98.0	59.2	97.7	85.4

Table 3. Confusion matrix for classification and separation of non-tree pollen types. Upper part: real classes vs. predicted classes of the classified test spectra; lower part: sensitivity, specificity, accuracy and precision in %.

difference, 784 nm and 786 nm data. The classification rates of non-tree types show even stronger variations between 85 and 98%. Here, the 784 and 786 nm data gave slightly better results. Overall, the classification results differed only little. Another comparison of baseline corrected Raman spectra with difference spectra collected by wavelength modulation was presented for leukocytes and tumor cells and confirmed only small difference in classification rates[57]. Similar to this pollen study, the spectral background was moderate in the previous cell study. A true benefit of SERDS is expected for high spectral background and contributions from ambient light and etaloning which cannot be suppressed by state-of-art baseline correction and goes well beyond.

Conclusions

In this contribution, the methodology was described to use difference spectra based on SERDS for classification of pollen data. Its main advantage is the analysis of single difference spectra by PCA and subsequent LDA with few PCs as input without complex data pre-processing. An optimization procedure compensated the photobleaching effects and minimized the remaining background in the difference spectra, whereas the normalization was necessary to obtain the same intensity range for all measured pollen difference spectra. This resulted in a classification based on the spectral features of the difference spectra and not based on the overall intensity of the Raman spectrum of a pollen sample. A further improvement would be rapid, serial acquisitions of Raman spectra at both wavelengths, which suppresses photobleaching effects and avoids the optimization procedure[58]. Since no reconstruction of the familiar profiles of Raman spectra is necessary, possible artefacts are not introduced into the spectra. The down side of this direct classification using SERDS spectra is the difficult spectral analysis of difference spectra and especially the resulting PC and LD loadings. The increase in noise level due to the subtraction of two spectra compared to a single Raman spectrum is compensated by PCA, which separated the spectral variations in the first PCs from the noise in the higher PCs. The possibility to classify different pollen samples using normalized and optimized difference spectra by a linear PCA-LDA model was successfully demonstrated. In Supplementary Table S1 the average values for sensitivity, specificity, accuracy and precision for all classifications are shown that achieved good to very good results using difference spectra as input. Since birch, hazel and alder all belong to the same family, in case of birch and alder even to the same subfamily, the pollen could even be separated on a genus level. For comparison, the reconstructed spectra and the raw spectra at 784 and 786 nm excitation after baseline correction were subjected to PCA-LDA classification. The classification rates only show small variations with a tendency of worse results for reconstructed spectra. This demonstrates the validity of our new approach based on difference spectra. The full potential of SERDS will become evident for Raman spectra that are affected by high autofluorescence background, ambient light or etaloning effects. Since the pollen detection, the laser focusing, the wavelength shifting and the data recording was fully automated, this streamlined method could be a robust and versatile system for the automated differentiation of different pollen into their genera.

Methods

Sample specifications. Pollen samples used in this study were collected by the Department of Indoor Climatology (University Hospital Jena, Germany) over the last two decades and transferred to the Leibniz Institute of Photonic Technology (Jena, Germany) for storage and research purposes. Eight different pollen genera were analyzed, which can be grouped into two classes based on their growth habit: tree and non-tree (see Supplementary Table S3).

Set-up description and specification. The previously developed high throughput screening Raman spectroscopy platform (HTS-RS)[59] was modified for the implementation of SERDS and the investigation of pollen grains. A tunable laser source (DLC DL pro 780, Toptica Photonics, Germany) with a tuning range from 765 to 805 nm in combination with an amplifier (BoosTA Pro, Toptica Photonics, Germany) was fiber-coupled into a microscope set-up by a multimode fiber with a 60 μm core diameter and 0.22 NA (Thorlabs, Germany).

The booster enhanced the laser power of the three operating wavelengths (λ = 784 nm/785 nm/786 nm) to the system. The incoming laser beam passed through a clean-up filter (785 ± 1.5 nm; Semrock, USA) and was collimated using a 30 mm focal length lens (Thorlabs, Germany). The collimated beam was guided to the back aperture of the microscope objective (60×, NA = 1, water immersion, Nikon, Japan) via a dichroic notch filter (785 nm, bandwidth 89 nm; Semrock, USA) and a 45° tilted mirror (Thorlabs, Germany). At the end of the objective, the pollen grains were excited with an approximate focus spot diameter of 10 µm on the sample plane. The back reflected Raman signals from the samples were collected and projected to a 100 mm focal length lens (Thorlabs, Germany) while passing through the same objective lens and the notch filter, which blocked most of the Rayleigh signal. An extra notch filter operating at 785 nm ± 19 nm (Laser Components, Germany) was placed before the 100 mm collection lens ensuring maximal rejection of Rayleigh signal propagation. The 100 mm collection lens focused the Raman signal to a 100 µm, 0.22 NA multimode fiber (Thorlabs, Germany) coupling it to a spectrograph (IsoPlane160, Princeton Instruments, USA) with a 400 grooves/mm grating blazed at 750 nm. The Raman signals were projected to a charge-coupled device (CCD) (PIXIS-400BR-eXcelon; Princeton Instruments, USA) with an operating temperature of − 60 °C. A bright field channel was integrated into the set-up for the automation of particle detection, of various calibrations and for visualization purposes. A white LED source (Thorlabs, Germany) was employed to illuminate the sample from below and the light was guided to a CCD camera (DCC1645C, Thorlabs, Germany) via a long pass filter (Semrock, USA) and a 70 mm focal length lens (Thorlabs, Germany). The bright field microscopic image was used in an in-house developed pollen detection algorithm for the automated detection of single pollen grain. All the required translations were realized using two CONEX MFA-Series motor (Newport, USA) for xy plane and a MTS25-Z8 motor (Thorlabs, Germany) for z direction[36,59].

Sample preparation and data acquisition. Each pollen sample was suspended in 10 mL deionized water and pipetted onto a CaF_2 cover slip fully immersed in deionized water. After sedimentation of the pollen, single pollen samples were automatically detected, the signal focused and measured at three different excitation wavelengths (λ_1 = 784 nm/130 mW; λ_2 = 785 nm/180 mW; λ_3 = 786 nm/200 mW) before moving to the next pollen. The acquisition time for each spectrum was 0.5 s with a short dwell time of 0.5 s after each measurement to allow for wavelength shifting.

Data preprocessing. The collected data was analyzed in R[60] using the following packages: hyperSpec[61], cbmodels[62], Ramancal[63], rgl[64], pracma[65], gtools[66] and ROCR[67]. The raw spectra were first corrected for cosmic spikes[68], wavelength calibrated in relation to λ_{ex} = 785 nm using 4-Acetaminophenole and then intensity calibrated using a white-light source calibration lamp (Raman Calibration Accessory—HCA, Kaiser Optical Systems, Inc., USA). Since the background between the different excitation wavelengths varies due to photobleaching and differences in laser intensity, the spectra of different excitation wavelengths were difference-optimized before subtraction as described previously[35]. For comparison, Raman spectra were reconstructed from the SERDS spectra by summation of the signal intensities ($S(n) = \sum_{x=1}^{n} S(x)$, where $S(n)$ corresponds to the signal intensity at pixel n) and subsequent baseline correction using the SNIP algorithm[14,69]. Furthermore, Raman spectra (λ_{ex} = 784 nm and λ_{ex} = 786 nm) were also baseline corrected (SNIP) after cosmic spike correction, wavelength and intensity calibration.

For each pollen type a Pearson correlation of the optimized difference spectra to the respective pollen type's mean optimized difference spectrum was performed as a spectral quality control to detect outliers. To make sure that also pollen debris spectra, out-of-focus pollen spectra and pure water spectra are excluded from the data set, all spectra with a Pearson correlation coefficient below 0.52 were discarded, leading to 4–24% of spectra being discarded. However, for the larch pollen the highest percentage of 52% was discarded because of poor sedimentation and therefore a lot of out-of-focus measurements due to the high lipid content of the pollen.

The overall intensity of the Raman signal varies between pollen types. To make sure the classification occurs due to spectral features and not overall intensity differences, the difference spectra were normalized. The performed normalization was a Euclidean-distance-like area normalization.

Data classification. The preprocessed, optimized and normalized data was separated into a test and training data set. The training data set was chosen in a way that every pollen type was represented by the same number of spectra (ca. 60% of the smallest data set resulting in 90 spectra). The remaining spectra were used for testing. As a classification scheme the training data was first dimensionally reduced using a Principal Component Analysis (PCA) and then used to construct a Linear Discriminant Analysis (LDA) classification model. The models were internally validated using a tenfold cross validation to determine the optimal number of principal components (PCs) for building robust PCA-LDA models. In a first step 10 PCs of all eight pollen genera were classified into the two growth habits: tree and non-tree. In a second step 13 PCs of all measured tree pollen were classified based on their genera: alder, birch, hazel and larch. In the final step 11 PCs of all measured non-tree pollen were classified into their different genera: mugwort, cyclamen, moor grass and rumex. Herein, we present the $\Delta\lambda_{ex}$ = 2 nm shift in excitation wavelength (λ_1 = 784 nm; λ_2 = 786 nm; $\Delta\lambda_{ex} = \lambda_1 - \lambda_2$ = 2 nm). For comparison, the same classification schemes were used on the reconstructed Raman spectra and the baseline corrected Raman spectra.

References

1. Germond, A. *et al.* Raman spectroscopy as a tool for ecology and evolution. *J. R. Soc. Interface* **14**, 20170174 (2017).
2. Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **11**, 664–687 (2016).
3. Cheng, J. X. & Xie, X. S. Vibrational spectroscopic imaging of living systems: An emerging platform for biology and medicine. *Science* **350**, aaa8870 (2015).
4. Hubbard, T. J. E., Shore, A. & Stone, N. Raman spectroscopy for rapid intra-operative margin analysis of surgically excised tumour specimens. *Analyst* **144**, 6479–6496 (2019).
5. Monici, M. Cell and tissue autofluorescence research and diagnostic applications. *Biotechnol. Annu. Rev.* **11**, 227–256 (2005).
6. Zięba-Palus, J. & Michalska, A. Photobleaching as a useful technique in reducing of fluorescence in Raman spectra of blue automobile paint samples. *Vib. Spectrosc.* **74**, 6–12 (2014).
7. Yakubovskaya, E., Zaliznyak, T., Martínez, J. & Taylor, G. T. Tear down the fluorescent curtain: A new fluorescence suppression method for raman microspectroscopic analyses. *Sci. Rep.* **9**, 1–9 (2019).
8. Wei, D., Chen, S. & Liu, Q. Review of fluorescence suppression techniques in Raman spectroscopy. *Appl. Spectrosc. Rev.* **50**, 387–406 (2015).
9. Afseth, N. K. & Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **117**, 92–99 (2012).
10. Martens, H. & Stark, E. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *J. Pharm. Biomed. Anal.* **9**, 625–635 (1991).
11. Stark, E. W. & Martens, H. Multiplicative signal correction method and apparatus. US Patent US005568400A, 1–19 (1990).
12. Pirzer, M. & Sawatzki, J. Patent Application Publication Pub. No.: US 2006/0211562 A1. 1–11 (2006).
13. Kneen, M. A. & Annegarn, H. J. Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **109–110**, 209–213 (1996).
14. Morhac, M. Software Package for R—Peaks: Background estimation, Markov smoothing, deconvolution and peaks search functions. (2012). https://rdrr.io/cran/Peaks/.
15. Mahadevan-Jansen, A. & Lieber, C. A. Automated method for subtraction of fluorescence from biological Raman spectra. *Appl. Spectrosc.* **57**, 1363–1367 (2003).
16. Cordero, E. *et al.* Bladder tissue characterization using probe-based Raman spectroscopy: Evaluation of tissue heterogeneity and influence on the model prediction. *J. Biophotonics* **13**, e201960025 (2020).
17. Bergholt, M. S. *et al.* Characterizing variability of in vivo Raman spectroscopic properties of different anatomical sites of normal colorectal tissue towards cancer diagnosis at colonoscopy. *Anal. Chem.* **87**, 960–966 (2015).
18. Desroches, J. *et al.* A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy. *Sci. Rep.* **8**, 1–10 (2018).
19. Galli, R. *et al.* Rapid label-free analysis of brain tumor biopsies by near infrared Raman and fluorescence spectroscopy—A study of 209 patients. *Front. Oncol.* **9**, 1–13 (2019).
20. Shreve, A. P., Cherepy, N. J. & Mathies, R. A. Effective rejection of fluorescence interference in Raman spectroscopy using a shifted excitation difference technique. *Appl. Spectrosc.* **46**, 707–711 (1992).
21. Dochow, S. *et al.* Etaloning, fluorescence and ambient light suppression by modulated wavelength Raman spectroscopy. *Biomed. Spectrosc. Imaging* **1**, 383–389 (2012).
22. Maiwald, M. *et al.* Microsystem 671 nm light source for shifted excitation Raman difference spectroscopy. *Appl. Opt.* **48**, 2789 (2009).
23. Bell, S. E. J. J., Bourguignon, E. S. O. O. & Dennis, A. Analysis of luminescent samples using subtracted shifted Raman spectroscopy. *Analyst* **123**, 1729–1734 (1998).
24. Maiwald, M. *et al.* Microsystem light source at 488 nm for shifted excitation resonance Raman difference spectroscopy. *Appl. Spectrosc.* **63**, 1283–1287 (2009).
25. Kiefer, J. Instantaneous shifted-excitation Raman difference spectroscopy (iSERDS). *J. Raman Spectrosc.* **45**, 980–983 (2014).
26. Sowoidnich, K. & Kronfeldt, H.-D. Fluorescence rejection by shifted excitation Raman difference spectroscopy at multiple wavelengths for the investigation of biological samples. *ISRN Spectrosc.* **2012**, 1–11 (2012).
27. Noack, K. *et al.* Combined shifted-excitation Raman difference spectroscopy and support vector regression for monitoring the algal production of complex polysaccharides. *Analyst* **138**, 5639–5646 (2013).
28. Martins, M. A. *et al.* Shifted-excitation Raman difference spectroscopy for in vitro and in vivo biological samples analysis. *Biomed. Opt. Express* **1**, 617 (2010).
29. Gebrekidan, M. T. *et al.* A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference. *J. Raman Spectrosc.* **47**, 198–209 (2016).
30. Maiwald, M., Müller, A., Sumpf, B., Erbert, G. & Tränkle, G. Capability of shifted excitation Raman difference spectroscopy under ambient daylight. *Appl. Opt.* **54**, 5520 (2015).
31. Maiwald, M., Müller, A., Sumpf, B. & Tränkle, G. A portable shifted excitation Raman difference spectroscopy system: Device and field demonstration. *J. Raman Spectrosc.* **47**, 1180–1184 (2016).
32. Schmälzlin, E. *et al.* Ultrafast imaging Raman spectroscopy of large-area samples without stepwise scanning. *J. Sens. Sens. Syst.* **5**, 261–271 (2016).
33. Zhao, J., Carrabba, M. M. & Allen, F. S. Automated fluorescence rejection using shifted excitation Raman difference spectroscopy. *Appl. Spectrosc.* **56**, 834–845 (2002).
34. Guo, S., Chernavskaia, O., Popp, J. & Bocklitz, T. Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS). *Talanta* **186**, 372–380 (2018).
35. Cordero, E. *et al.* Evaluation of shifted excitation raman difference spectroscopy and comparison to computational background correction methods applied to biochemical Raman spectra. *Sensors (Switzerland)* **17**, 1724 (2017).
36. Mondol, *et al.* Application of high-throughput screening Raman spectroscopy (HTS-RS) for label-free identification and molecular characterization of pollen. *Sensors* **19**, 4428 (2019).
37. Holt, K., Allen, G., Hodgson, R., Marsland, S. & Flenley, J. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Rev. Palaeobot. Palynol.* **167**, 175–183 (2011).
38. Haas, N. Q. *Automated Pollen Image Classification.* Master thesis, University of Tennessee (2011).
39. Koutsoukos, I. *Automated Classification of Pollen Grains from Microscope Images using Computer Vision and Semantic Web Technologies.* Diploma thesis, Technical University of Crete (2013).
40. Dell'Anna, R. *et al.* Pollen discrimination and classification by Fourier transform infrared (FT-IR) microspectroscopy and machine learning. *Anal. Bioanal. Chem.* **394**, 1443–1452 (2009).
41. Gottardini, E., Rossi, S., Cristofolini, F. & Benedetti, L. Use of Fourier transform infrared (FT-IR) spectroscopy as a tool for pollen identification. *Aerobiologia (Bologna)* **23**, 211–219 (2007).
42. Pappas, C. S., Tarantilis, P. A., Harizanis, P. C. & Polissiou, M. G. New method for pollen identification by FT-IR spectroscopy. *Appl. Spectrosc.* **57**, 23–27 (2003).
43. Samuels, A. C., DeLucia, F. C., McNesby, K. L. & Miziolek, A. W. Laser-induced breakdown spectroscopy of bacterial spores, molds, pollens, and protein: Initial studies of discrimination potential. *Appl. Opt.* **42**, 6205 (2003).

44. Schulte, F., Panne, U. & Kneipp, J. Molecular changes during pollen germination can be monitored by Raman microspectroscopy. *J. Biophotonics* **3**, 542–547 (2010).
45. Laucks, M. L., Roll, G., Schweiger, G. & Davis, E. J. Physical and chemical (RAMAN) characterization of bioaerosols-pollen. *J. Aerosol. Sci.* **31**, 307–319 (2000).
46. Bağcıoğlu, M., Zimmermann, B. & Kohler, A. A multiscale vibrational spectroscopic approach for identification and biochemical characterization of pollen. *PLoS ONE* **10**, 1–19 (2015).
47. Merlin, J. C. Resonance Raman spectroscopy of carotenoids and carotenoid-containing systems. *Pure Appl. Chem.* **57**, 785–792 (2007).
48. Wang, C., Pan, Y. L., Hill, S. C. & Redding, B. Photophoretic trapping-Raman spectroscopy for single pollens and fungal spores trapped in air. *J. Quant. Spectrosc. Radiat. Transf.* **153**, 4–12 (2015).
49. Pummer, B. G. *et al.* Chemistry and morphology of dried-up pollen suspension residues. *J. Raman Spectrosc.* **44**, 1654–1658 (2013).
50. Seifert, S., Merk, V. & Kneipp, J. Identification of aqueous pollen extracts using surface enhanced Raman scattering (SERS) and pattern recognition methods. *J. Biophotonics* **9**, 181–189 (2016).
51. Schulte, F., Mäder, J., Kroh, L. W., Panne, U. & Kneipp, J. Characterization of pollen carotenoids with in situ and high-performance thin-layer chromatography supported resonant Raman spectroscopy. *Anal. Chem.* **81**, 8426–8433 (2009).
52. Zimmermann, B. Characterization of pollen by vibrational spectroscopy. *Appl. Spectrosc.* **64**, 1364–1373 (2010).
53. Schulz, H., Baranska, M. & Baranski, R. Potential of NIR-FT-Raman spectroscopy in natural carotenoid analysis. *Biopolymers* **77**, 212–221 (2005).
54. Sengupta, A., Laucks, M. L. & James Davis, E. Surface-enhanced Raman spectroscopy of bacteria and pollen. *Appl. Spectrosc.* **59**, 1016–1023 (2005).
55. Schulte, F., Lingott, J., Panne, U. & Kneipp, J. Chemical characterization and classification of pollen. *Anal. Chem.* **80**, 9551–9556 (2008).
56. Ivleva, N. P., Niessner, R. & Panne, U. Characterization and discrimination of pollen by Raman microscopy. *Anal. Bioanal. Chem.* **381**, 261–267 (2005).
57. Dochow, S. *et al.* Classification of Raman spectra of single cells with autofluorescence suppression by wavelength modulated excitation. *Anal. Methods* **5**, 4608–4614 (2013).
58. Sowoidnich, K., Towrie, M., Maiwald, M., Sumpf, B. & Matousek, P. Shifted excitation Raman difference spectroscopy with charge-shifting CCD lock-in detection. *Appl. Spectrosc.* **73**, 1265–1276 (2019).
59. Schie, I. W. *et al.* High-throughput screening Raman spectroscopy platform for label-free cellomics. *Anal. Chem.* **90**, 2023–2030 (2018).
60. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
61. Beleites, C. & Sergo, V. Software Package for R-hyperSpec: A package to handle hyperspectral data sets in R. (2018).
62. Beleites, C. Software Package for R-cbmodels: Collection of 'combined' models: PCA-LDA, PLS-LDA, PLS-LR as well as EMSC. (2015).
63. Beleites, C. Software Package for R-Ramancal: Calibration routines for Raman spectrometers. (2013).
64. Adler, D., Murdoch, D. & others. Software Package for R-rgl: 3D Visualization Using OpenGL. (2018).
65. Borchers, H. W. Software Package for R-pracma: Practical Numerical Math Functions. (2018).
66. Warnes, G. R., Bolker, B. & Lumley, T. Software Package for R-gtools: Various R Programming Tools. (2018).
67. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. Software Package for R-ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**, 7881 (2005).
68. Ryabchykov, O. Software Package for R-Spikes: Spike Correction of Raman Spectral Data. (2016).
69. Gibb, S. & Strimmer, K. MALDIquant: A versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28**, 2270–2271 (2012).

Acknowledgements

Author contributions

The initial conceptualization and methodology was done by F.K., C.S., C.K. and J.P. A.S.M. and I.W.S. built the set-up and programmed the automated pollen detection, automated focus and automated shifting of the laser source. F.K. performed the measurements and the data analysis. C.K., I.W.S. and J.P. provided materials, equipment and laboratory. All authors discussed and agreed the results. F.K. and A.S.M. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. Supervision of the work was done by C.S., C.K. and J.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67897-4.

Correspondence and requests for materials should be addressed to C.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.