

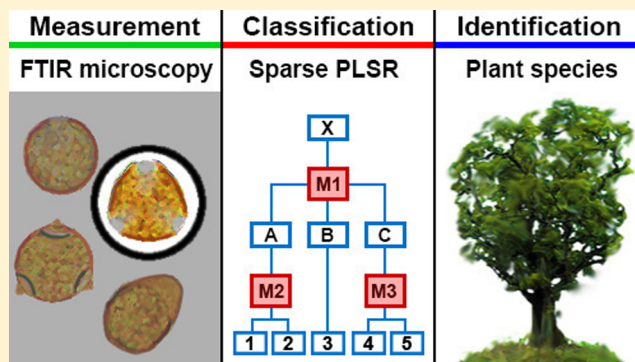
Analysis of Allergenic Pollen by FTIR Microspectroscopy

B. Zimmerman,* V. Tafintseva, M. Bağcıoğlu, M. Høegh Berdahl, and A. Kohler

Department of Mathematical Sciences and Technology, Faculty of Environmental Science and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway

S Supporting Information

ABSTRACT: Fourier transform infrared (FTIR) spectroscopy is a powerful tool for the identification and characterization of pollen and spores. However, interpretation and multivariate analysis of infrared microscopy spectra of single pollen grains are hampered by Mie-type scattering. In this paper, we introduce a novel sampling setup for infrared microspectroscopy of pollens preventing strong Mie-type scattering. Pollen samples were embedded in a soft paraffin layer between two sheets of polyethylene foils without any further sample pretreatment. Single-grain infrared spectra of 13 different pollen samples, belonging to 11 species, were obtained and analyzed by the new approach and classified by sparse partial least-squares regression (PLSR). For the classification, chemical and physical information were separated by extended multiplicative signal correction and used together to build a classification model. A training set of 260 spectra and an independent test set of 130 spectra were used. Robust sparse classification models allowing the biochemical interpretation of the classification were obtained by the sparse PLSR, because only a subset of variables was retained for the analysis. With accuracy values of 95% and 98%, for the independent test set and full cross-validation respectively, the method is outperforming the previously published studies on development of an automated pollen analysis. Since the method is compatible with standard air-samplers, it can be employed with minimal modification in regular aerobiology studies. When compared with optical microscopy, which is the benchmark method in pollen analysis, the infrared microspectroscopy method offers better taxonomic resolution, as well as faster, more economical, and bias-free measurement.



Allergenic pollens and spores trigger asthma, allergic rhinitis, and hypersensitivity pneumonitis in susceptible individuals. These respiratory diseases are affecting 10–20% of European population, causing multibillion costs of medical treatments and reduced work productivity.^{1,2} Moreover, the prevalence of respiratory allergic diseases has increased markedly during the last 20 years.³ It is expected that shifts in aeroallergen human exposures due to global climate change are going to increase the prevalence and severity of allergies.^{4,5} The risks analysis of urban atmospheric pollution in Madrid has showed that the effect of allergenic pollen on the health of individuals is greater than the detected effect of traditionally analyzed parameters, such as chemical pollution.⁶ Qualitative and quantitative analysis of pollen is the cornerstone of aeroallergen monitoring networks. Providing information on the occurrence of aeroallergens in the residential and working areas is vital for timely treatment of inhalant allergies through correlation of symptoms. Moreover, air quality forecasts are essential for avoidance and prevention that reduces allergic symptoms. Finally, palynology, the study of pollen, spores, and similar bioparticles, has great general interest in analysis of pollen grains, for example, in studies of environmental and climate change, biodiversity monitoring, and tracing vegetation history.

Current identification of pollen relies on optical (transmitted light) microscopy. Identification is based on qualitative descriptions of pollen's morphology, usually size, shape, and texture, as well on its substructures. The analysis is time-intensive, requiring extensive sample preparation and laborious identification by experienced specialists. The identification is limited to categories that can be reliably classified by multiple analysts, resulting in the rough taxonomic resolution. In addition, microscopy methods are prone to inconsistency and a degree of subjectivity because the performance of the individual analyst can arise through issues like fatigue and overfamiliarity with samples. For this reason, there is a trend in moving toward computer-based systems that deliver faster, more accurate, and more consistent semi- or fully automated identifications than any human operator.^{7,8} However, even if traditional microscopies increase sample processing by automation, the morphological approach is limited since many species have indistinguishable morphology. For instance, grass family (Poaceae) is one of the largest plant families with over 10 000 species, but grass pollens are rarely identified at the

Received: August 20, 2015

Accepted: November 24, 2015

Published: November 24, 2015

subfamily level. Alternative analyses, such as genetic sequencing and biochemical methods, are even more time-consuming, expensive, and difficult to implement in routine analysis.

Infrared spectroscopy is a biophysical method that provides precise signatures of the overall biochemical composition of pollen. Infrared spectra of pollen contain specific signals of lipids, proteins, carbohydrates, and cell wall biopolymers, such as cellulose and sporopollenins. Therefore, Fourier transform infrared (FTIR) spectroscopy has emerged as a powerful tool in identification and characterization of pollen and spores.^{9–22} Moreover, FTIR spectroscopy offers an operator-independent approach based on chemical fingerprinting via identifiable spectral features, with significant potential for automation of pollen identification.⁷ Since samples can be measured in their native form, without any chemical pretreatment, FTIR spectroscopy achieves simple, economical and rapid analysis of pollen. In FTIR microspectroscopy (μ FTIR), an infrared spectrometer is combined with a microscope, thus enabling focused measurement of an individual bioparticle. Unfortunately, bioparticles such as pollen are highly scattering particles in the infrared due to their size and morphology. Therefore, strong scattering signatures in μ FTIR spectra of single pollen grains hamper biochemical interpretation of spectra.^{15,22,23} The strong scattering by the measured pollen grains results in anomalous spectral features that significantly interfere with and distort the signals of chemical absorption.²³ Extracting information from recorded data by separating absorption and scattering contributions is extremely challenging because scattering and absorption are highly entangled. Moreover, the pollen shapes and structures differ considerably from spherical and homogeneous particles as assumed for existing correction models.^{23,24} In general, scattering effects have been identified as a major obstacle for application of FTIR microspectroscopy in biological and biomedical science.

In the presented study, we have implemented an innovative approach for scatter-free FTIR microscopic measurement of single bioparticles, specifically single pollen grains. The aim of the study was to test the new approach for aeroallergenic pollen species by spectroscopy-based identification of plant taxa. The chemical characterization of pollen grains and extraction of morphological parameters from spectral data is discussed in details. Moreover, the complementarity of FTIR microspectroscopy and contemporary air-sampler systems has been taken into consideration. The presented approach would enable rapid implementation of FTIR microspectroscopy in aeroallergen monitoring networks, as well as in environmental and climate change studies.

■ EXPERIMENTAL SECTION

Samples. The pollen types used were collected from *Fagus sylvatica* (European beech), *Betula pendula* (Silver birch), *Carpinus betulus* (European hornbeam), *Quercus cerris* (Turkey oak) and *Quercus robur* (Penduculate oak) of the Fagales (beech) order, *Bromus erectus* (Erect brome) and *Carex pendula* (Pendulous sedge) of the Poales (grass) order, *Cryptomeria japonica* (Japanese cedar), *Cupressus sempervirens* (Mediterranean cypress), *Juniperus phoenicea* (Phoenicean juniper) and *Juniperus chinensis* (Chinese juniper) of the Pinales (pine) order. *Betula pendula* was represented with three samples, collected from three different individuals: one wild type and two “Youngii” cultivars. All other species were represented with a single sample. Each sample comprised pollen belonging to either: (1) approximately 10 inflorescences from the same turf

(for Poales), (2) approximately 10 male catkins from the same individual plant (for Fagales), and (3) approximately 50 male strobili from the same individual plant (for Pinales). In total, 13 pollen samples were measured. Pollen samples were obtained through fieldwork at the Botanical Garden of the Faculty of Science, University of Zagreb, during 2012 pollination seasons. The pollen samples were collected directly from plants at flowering time, and stored at $-15\text{ }^{\circ}\text{C}$.

Measurement. The pollen samples were embedded, without any sample pretreatment, in a matrix medium, consisting of a soft paraffin (Vaseline-type petroleum jelly) between two sheets of $40\text{ }\mu\text{m}$ thick polyethylene foils. The amount of paraffin was approximately $1\text{ mg per }25\text{ cm}^2$. A custom-made sample holder was made in order to stretch the foils and to minimize the appearance of air bubbles in the sample (Figure S-1, in the Supporting Information).

Infrared spectroscopy measurements were obtained on a VERTEX 70 FTIR spectrometer with a HYPERION 3000 IR microscope (Bruker Optik GmbH, Germany). The spectra were recorded in transmission mode with a spectral resolution of 4 cm^{-1} and digital spacing of 1.929 cm^{-1} . Thirty single grains were measured per pollen sample, and for each grain, one corresponding FTIR spectrum was obtained. The system is equipped with a global mid-IR source and a liquid nitrogen-cooled mercury cadmium telluride (MCT) detector. The spectra were measured in the $4000\text{--}600\text{ cm}^{-1}$ spectral range, with 64 scans for both background and sample spectra, and using a $15\times$ objective, with $25\times 25\text{ }\mu\text{m}$ aperture sizes. For the measurements in the matrix media, background (reference) spectra of air were recorded once, before starting each (sample) measurement using the sample-free setup. The quality of recorded spectra was measured by the performing Bruker OPUS quality test for measurement of microorganisms, by calculating the signal-to-noise ratio of Amide I signal (at $1700\text{--}1600\text{ cm}^{-1}$).

The referent (matrix-free) single grain pollen spectra were measured by depositing pollen samples onto 3 mm ZnSe optical slides (windows) without any sample pretreatment. For the measurements on ZnSe optical windows, background (reference) spectra were recorded immediately before starting each measurement by measuring empty areas of ZnSe windows. The referent bulk pollen spectra were recorded by measuring potassium bromide (KBr) sample pellets. KBr pellets were prepared by mixing and grinding approximately 1 mg of a sample with approximately 100 mg of KBr using a pestle and mortar, and the created KBr matrix was then cold-pressed without degassing into a transparent disk. A sample-free setup was used to obtain the spectral background.

The Bruker system was controlled with OPUS 7.5 software (Bruker Optik GmbH, Germany).

Spectral Preprocessing. From each of the 13 pollen samples, 30 single grain spectra of 30 different pollen grains were obtained. In total, 360 μ FTIR spectra were included in the data analysis. The spectra were preprocessed prior to calibration. The spectral region of 1900 cm^{-1} to 900 cm^{-1} was selected for data analysis, excluding the region of 1520 to 1290 cm^{-1} , which is the saturation region of the polyethylene foil and paraffin (i.e., the region with strong absorbance bands of matrix media). Two types of preprocessing procedures were used in the study: (1) preprocessing of second-derivative spectra and (2) preprocessing of nonderivated spectra. For the derivated data, all spectra were transformed to second-derivative spectra by the Savitzky–Golay algorithm using a

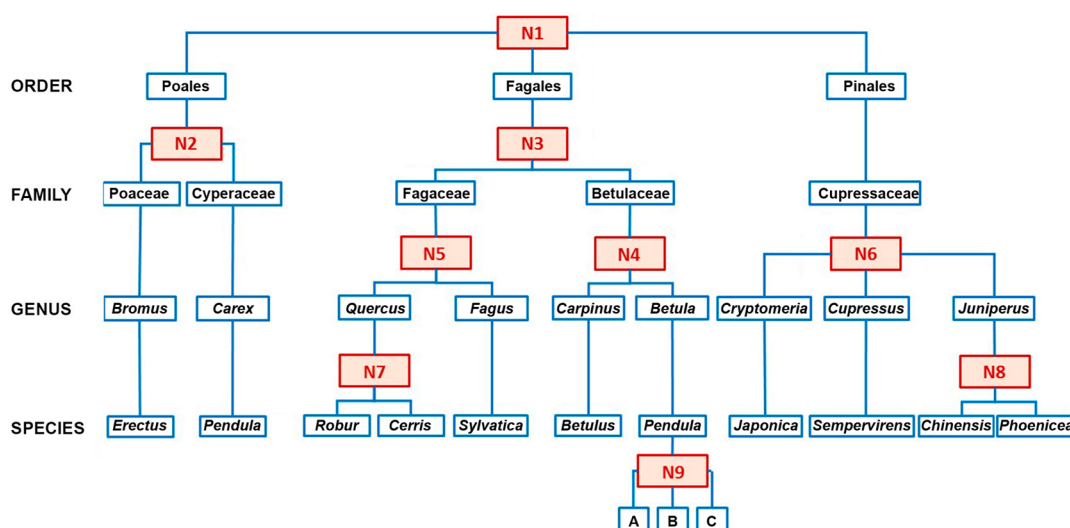


Figure 1. Phylogenetic tree for classification of pollen spectra, containing nine nodes (N1–N9).

polynomial of degree two and a window size of 31 points in total, followed by extended multiplicative signal correction (EMSC, an MSC model extended by a linear, quadratic and cubic component).^{25–27} The selection of window size was based on suppressing noise while enhancing intensity of broad amide bands in the 1650–1520 cm^{-1} region, as explained in detail in our previous study.²⁶ The EMSC model used in the preprocessing is defined by equations S-1 and S-2, in the [Supporting Information](#).

For the nonderived data set, all spectra were preprocessed by EMSC. In the EMSC preprocessing of nonderived data, the spectral region of chemical absorbance was down-weighted, and spectral regions devoid of any chemical absorbance were up-weighted, by applying a weighting vector. Vector value 1 was used in the whole spectral region, except the region 1900–1800 cm^{-1} , where the weighting vector was set to 10. The region 1900–1800 cm^{-1} is devoid of any chemical absorbance and thus this region should have the same baseline values in all pollen spectra when physical effects (such as light reflection) have been removed. Up-weighting of this region, by applying a weighting vector, is constraining the EMSC preprocessing and ensuring a stable baseline in all pollen spectra. For the data analysis, three different data sets were used: (A) the derived and EMSC preprocessed spectral data set alone, (B) the EMSC parameters data set obtained from the nonderived spectra, and (C) a combined data set where EMSC parameters obtained from the nonderived spectra were concatenated with the derived and EMSC preprocessed spectra. The spectral preprocessing was performed by algorithms developed in-house in the environment of Matlab, R2014a (The Mathworks Inc., Natick, United States).

Data Analysis and Classification. The data sets A and C were used to evaluate classification and biochemical similarities between pollen samples by using principal component analysis (PCA) and sparse partial least-squares regression (Sparse PLSR). Sparse PLSR was used for building a hierarchical classification scheme. The hierarchical classification scheme was built following the phylogenetic tree shown in [Figure 1](#) allowing classification at various subcategories such as order, family, genus, species, and sample. In each node (N1–9 in [Figure 1](#)) a classification model was established for 2 or 3 classes, depending on the number of subclasses connected to each

node. For establishing the classification models, matrices of indicator variables were set up for each node (Y-matrices). The indicator matrices contain one column for each subclass of the nodes with numbers ones and zeros indicating if a sample belongs to a respective subclass or not. Sparse PLSR was used to regress the spectral data sets (X-data sets) on the matrices of indicator variables (Y-matrices) at each node. In analogy to discriminant PLSR,²⁸ we denote the sparse PLSR model used for classification as discriminant Sparse PLSR. Sparse PLSR allows straightforward biochemical interpretation of spectral bands because only a limited subset of variables is selected in the modeling process.²⁹ Three classification models, using two different data sets (as defined in the [Spectral Preprocessing](#) section), were built on the basis of the tree shown in [Figure 1](#): 1) A sparse model based on data set A, 2) a sparse model based on data set C, and 3) a mixed sparse model based on data set A at all levels except “order” level (N1) where data set C was used. Each spectral data set (390 samples in total) was divided into a training set used for calibration (20 spectra per sample, 260 spectra in total), and an independent test set (10 spectra per sample, 130 spectra in total). The training set was used to establish Sparse PLSR classification models employing Venetian Blinds Cross-Validation (CV) with three segments where samples were split into blocks in the following order: 123 123.... For the training set, block separation was done by full or leave-one-out CV due to more reliable and faster performance. For each principal component, Sparse PLSR algorithm penalizes loading weight vectors according to a parameter called degree of sparsity. This parameter defines how many variables will be neglected and has previously been predefined.³⁰ In this paper, we use an optimization procedure for sparsity to choose the optimal degree of sparsity. For each PLSR component, an optimal degree of sparsity was chosen among four in the range from 80 to 99% of variables with a step of 4.75. The degree of sparsity, which gave the minimum missclassification rate (MCR) was selected. The selected degrees of sparsity were then used in the final PLSR calibration. Finally, the optimal number of components (AO_{pt}) was chosen. AO_{pt} was found by evaluating the MCR as a function of the number of components A. AO_{pt} is the one which does not have significantly higher MCR than the minimum MCR. A binomial distribution formed a basis for a binomial test, which

was used to evaluate the statistical significance (function `binocdf` in Matlab).

Data analyses were performed by algorithms developed in-house in the environment of Matlab, R2014a (The Mathworks Inc., Natick, United States).

RESULTS AND DISCUSSION

Scattering Spectral Artifacts and Experimental Setting. In Figure 2, spectra of single pollen grains measured on

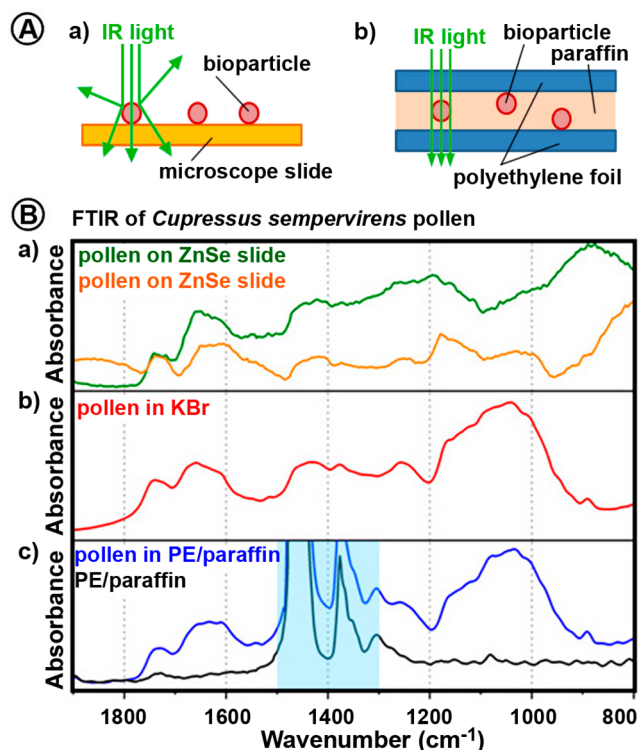


Figure 2. (A) Schematics of μ FTIR measurements of bioparticles on: (a) microscope slide and (b) within paraffin matrix; infrared light transmitted through and reflected on bioparticles is represented with arrows (specular reflection from microscope slide and polyethylene films is omitted for simplification). (B) Infrared spectra of the same *Cupressus sempervirens* pollen sample obtained with different experimental settings: (a) Spectra obtained by measurements of single pollen grains on microscope slides (unreproducible due to scatter effects); (b) Benchmark spectrum obtained by bulk measurement of a KBr sample pellet; (c) Scatter-free single grain spectrum (blue) obtained by measurement in paraffin and polyethylene (PE) embedding medium (spectrum of embedding media (black) is included for comparison). Blue shading designates the spectral region that was excluded from data analysis.

microscope slides are shown (Figure 2Ba). The spectra show substantial and complex anomalies when compared with the reference spectra of bulk samples (Figure 2Bb). Due to variations in morphological and optical properties of pollen grains within a sample, the corresponding pollen spectra have different scatter distortions. These spectral anomalies include wavelike fluctuating structures with narrower and broader signals, that can significantly interfere with and distort the signals of chemical absorption.^{22,23} It is worth noting that scatter contributions can provide valuable information on morphology of pollen grain. However, due to the high morphological variation of pollen grains within the same sample, they hinder classification of pollen types by FTIR

microspectroscopy of single pollen grains. In contrast, it is known that the biochemical composition as probed by FTIR spectroscopy is highly type-specific. Previous studies have shown that the biochemical information obtained by scatter-free absorbance spectra from KBr-pellet measurements of pollen bulk samples can separate pollen on different phylogenetic levels.²⁰ Thus, for the identification and chemical analysis of pollen, it is important to obtain pure (scatter-free) absorbance spectra that can serve as a fingerprint of the grain's biochemical components.

The scattering phenomena occurring in the μ FTIR spectroscopy of pollen have been interpreted as Mie-type scattering.²³ Bioparticles, such as pollen, are highly scattering samples in infrared spectroscopy because their sizes are often of the same order as the mid-infrared wavelengths employed in FTIR spectroscopy (2.5–25 μ m). This leads to strong Mie-type scattering.^{27,31–34} It is very difficult to separate the Mie-type scattering from the pure absorbance spectra. The reason is that the real refractive index undergoes strong fluctuations due to chemical absorbance bands and is thus highly dispersive. The real refractive index spectrum can be obtained from the pure absorbance spectrum, or equivalently from the imaginary part of the refractive index, by the Kramers–Kronig transform and vice versa. Unfortunately, for a measured infrared spectrum with Mie-type scattering, both the pure absorbance spectrum and the real refractive index spectrum are unknown. Recently, an iterative algorithm based on extended multiplicative signal correction (EMSC) for the retrieval of the pure absorbance spectra from infrared spectra with Mie-type scattering has been developed. Although the developed algorithm is robust and has provided some success, it remains unclear, to what extent, in general, it retrieves the pure absorbance spectrum, i.e. whether the algorithm converges toward the underlying pure absorbance spectrum or toward the reference spectrum used in the iterative EMSC approach. Another weakness of the algorithm is that it does not take into account the high numerical aperture of infrared microscopes. Recently, a different iterative algorithm has been presented taking into account the high numerical aperture.³⁵ This algorithm has been only tested on a model system (PMMA spheres) and could not be successfully applied for the correction of pollen.²³ Finally, all previous developed models for correcting Mie-type scattering in FTIR spectroscopy of bioparticles assume a spherical scatterer. While the shape of pollen grains is in many cases close to spheres, it can deviate considerably.¹⁴ In addition, the pollen structure is inhomogeneous, including a double layered grain wall and a multicellular grain interior made of cytoplasm, organelles, vacuoles, membranes, and nuclei.

In order to obtain scatter-free spectra of single pollen, we have developed an innovative experimental setting. Because the strength of the scatter-distortions depends on refractive index change at the surface of the bioparticle, scattering effects may be suppressed considerably by embedding samples into a medium with similar index of refraction as the bioparticle to be measured (Figure 2Ab). One suitable embedding matrix for μ FTIR of bioparticles is a thin layer of soft paraffin. In our study, 13 different plant samples were embedded in a paraffin matrix between two thin sheets of polyethylene foil and measured by μ FTIR. Thirty pollen grains per plant sample were measured, thus 360 spectra were obtained, each belonging to a different single grain. The obtained spectra are scatter-free (Figure 2Bc).

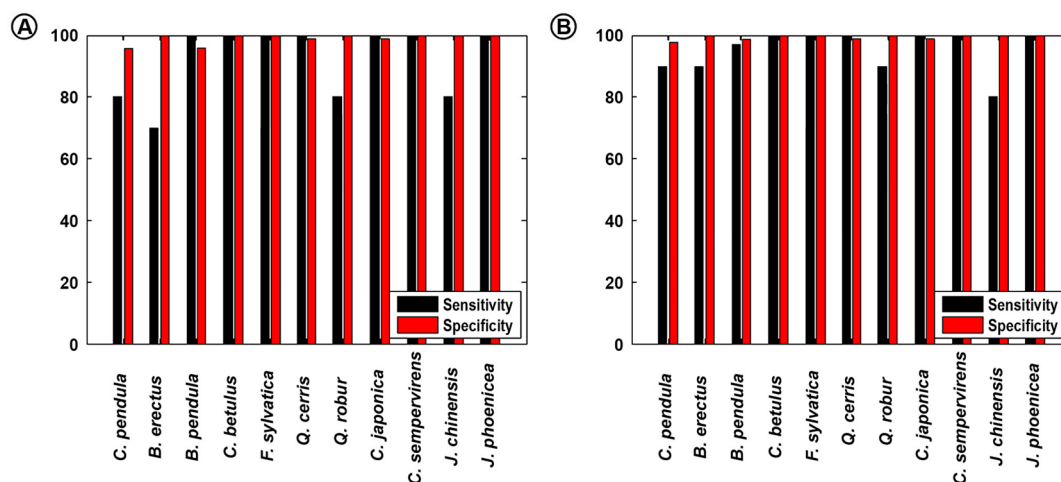


Figure 3. Identification results for Sparse PLSR analysis of pollen μ FTIR spectra, showing bar plots of sensitivity and specificity for each species (based on 10 spectra per sample). The results are based on (A) a sparse PLSR classification model built on second-derivative spectral data (data set A) and on (B) a sparse PLSR classification model built on second-derivative spectral data (data set A) at all levels except “order” level (N1) where a sparse PLSR classification model built on second-derivative spectral data with included EMSC parameters (data set C) was used.

The embedding matrix contains several strong absorption signals in the spectral region from 1500 to 1350 cm^{-1} . These bands are associated with CH_2 bending (1470 cm^{-1}), CH_3 antisymmetric (1463 cm^{-1}) and symmetric (1377 cm^{-1}) deformations, and CH_2 wagging vibrations (1353 and 1305 cm^{-1}).³⁶ The principal matrix-associated bands are over-saturated in all spectra, and therefore, the surrounding region, from 1520 to 1290 cm^{-1} , was excluded from the analysis (Figure 2Bc). Because this region is devoid of principal absorption signals of pollen, the subsequent spectral data analysis was not affected by large information shortfall. Additional information on sample measurement, including Figure S-2, are provided in the [Supporting Information](#).

Classification and Discrimination of Pollen. Pollen grains have unique vibrational spectra. PCA plots (Figure S-3, in the [Supporting Information](#)) show well-defined clustering at lower taxonomic levels, including species level. By using the first four principal components, spectra of samples belonging to distinct plant species can be separated quite well from each other. Moreover, EMSC parameter estimates of pollen μ FTIR spectra have a range of taxon-specific values (Figures S-4 and S-5, in the [Supporting Information](#)). Although EMSC parameters, containing baseline, background, and other anomalous features, are often discarded in spectral analysis, they may contain information that is important for characterization of pollen.²¹ EMSC parameter information is closely related to a characteristic morphology of a pollen grain, and pollen morphology varies substantially in shape, size, and wall pattern across plant taxa. As such, EMSC parameters offer an indirect physical measurement of a pollen sample that should be used for pollen characterization, alongside chemical absorption spectrum. Therefore, μ FTIR spectra of pollen show large potential for classification analysis.

Multivariate methods such as PCA and PLSR, allow investigating to what extent infrared spectra can be used for classification and discrimination of pollen types. PCA and PLSR are based on latent variables (also called scores) and permit direct interpretation of spectral signatures on the biochemical level. This interpretation is based on assessing regression coefficients and loading vectors. The regression coefficients relate the FTIR variables (wavelengths and thus

spectral bands) directly to the classes, while loading vectors relate the new latent variables to the FTIR variables. Evaluation of biochemical fingerprints is important, because a meaningful biological interpretation of the regression coefficients represents an additional validation of the discrimination and classification between different pollen on taxonomical level, for example differentiation of families, genera, species, or populations. However, interpretation of results can be challenging in situations where the number of variables is high, such is the case for FTIR data. In such cases, correlation coefficients and loading vectors are too complex for interpretation as a result of the multitude of spectral signatures revealed. Sparse PLSR combines variable selection and modeling, and thus enhances the interpretation of the classification analysis by discarding uninformative spectral variables.³⁰ The Sparse PLSR classification of the measured pollen spectra was based on the phylogenetic tree presented in Figure 1. As can be seen, the tree contained nine nodes (N1–N9) that split at different levels (order, family, genus, species, and sample) to different numbers of subclasses (two or three). Thus, the Sparse PLSR method was used to establish 9 calibration models, one for each node in the tree.

Among the 9 calibration models, the most complex model included AO_{opt} with 4 and 6 components, for data set A and data set C, respectively, with average complexity of $AO_{\text{opt}} = 2$ for both data sets. The advantage of sparse PLSR modeling lies in easier interpretation and often lower complexity of the models.

Sensitivity and specificity of discrimination analyses are presented in Figure 3. When considering only chemical spectral data (data set A), overall accuracy was 93.1%, with all except one taxa having sensitivity greater than or equal to 80% (Figure 3A). When EMSC parameters were included in the analysis (data set C), overall accuracy remained at 93.1%; however, all taxa had sensitivity equal or better than 80% (Figure S-6, in the [Supporting Information](#)). Specificity was high (more than 95%) in all cases. Although both data sets obtain the same overall accuracy, they differ substantially regarding accuracy at “order” level of classification (node N1 in Figure 1) where data set C performs much better than data set A. When EMSC parameters were included in the analysis (data set C), misclassification of

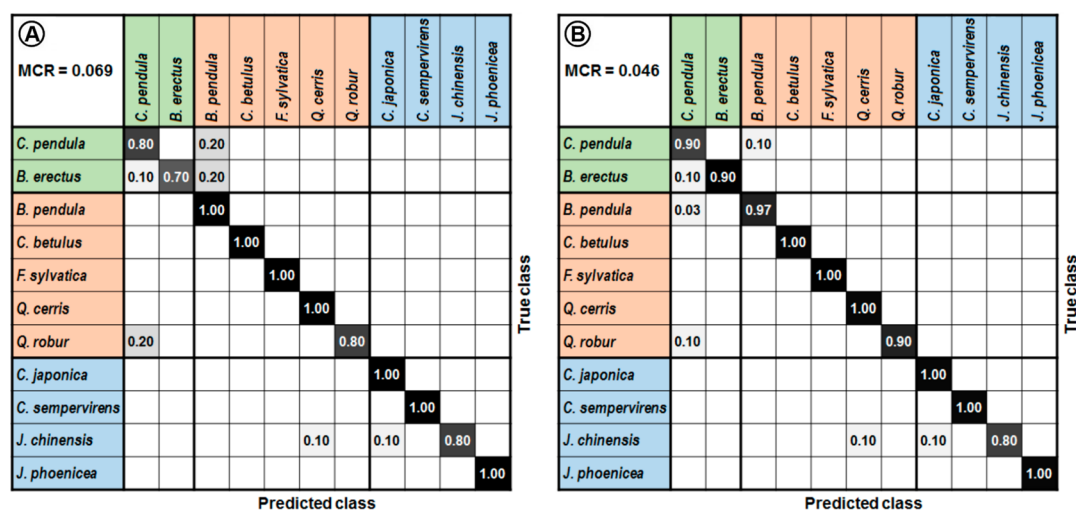


Figure 4. Confusion matrix for Sparse PLSR analysis of pollen μ FTIR spectra; results based on (A) a sparse PLSR classification model built on second-derivative spectral data (data set A) and on (B) a sparse PLSR classification model built on second-derivative spectral data (data set A) at all levels except “order” level (N1) where a sparse PLSR classification model built on second-derivative spectral data with included EMSC parameters (data set C) was used. Color codes: green (Poales), red (Fagales), and blue (Pinales). (MCR = misclassification rate).

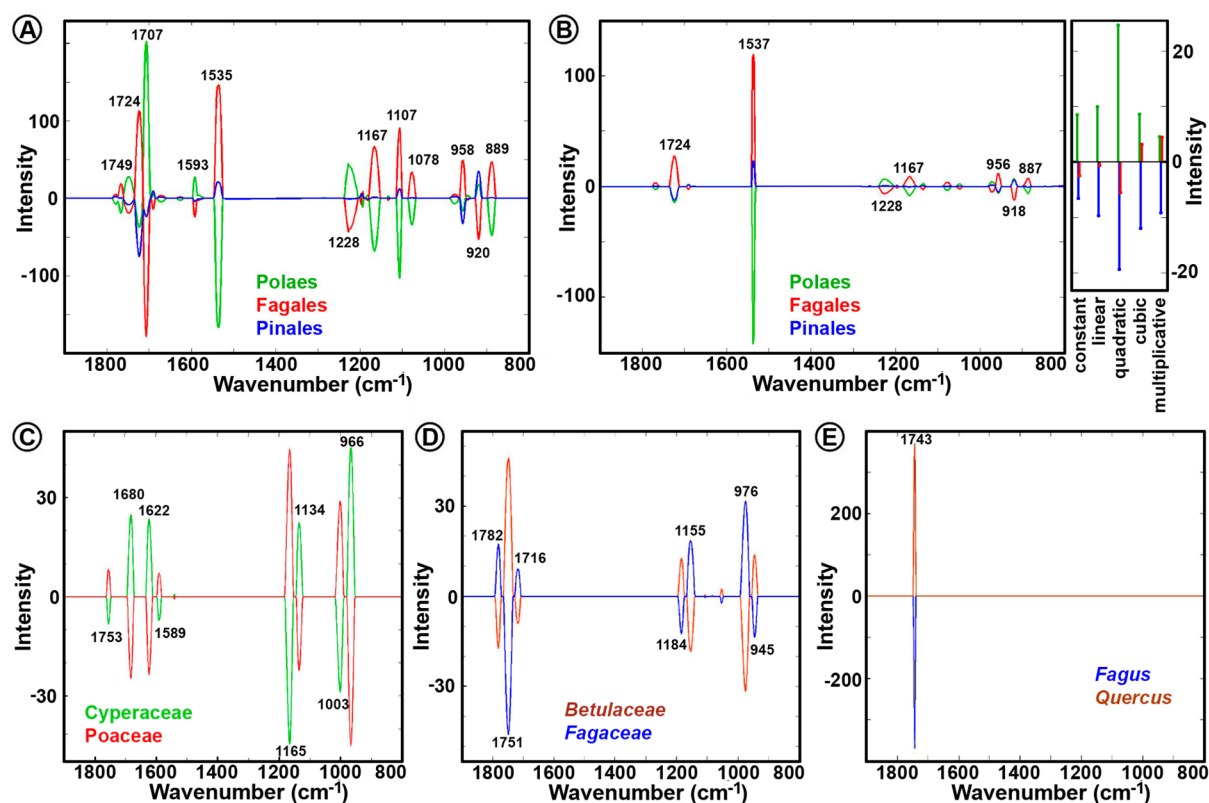


Figure 5. Plot of Sparse PLSR regression coefficients for second-derivative spectral data: classification of plant orders (node N1) by using (A) second-derivative spectral data (data set A) and (B) second-derivative spectral data with included EMSC parameters (data set C); classification by using second-derivative spectral data (data set A) of (C) Poales families (node N2), (D) Fagales families (node N3), and (E) Fagaceae genera (node N5).

plant orders was reduced by 43%. Therefore, we introduced a classification model combining a sparse PLSR classification model built on data set A and a sparse PLSR classification model built on data set C. We used a sparse PLSR classification model built on data set A at all levels except “order” level where a sparse PLSR classification model based on data set C was used. The model has resulted with an overall accuracy of 95.4%

and specificity equal or more than 98% for all species (Figure 3B).

Figure 4 presents the corresponding confusion matrices for the analyzed data sets with independent validation. In general, inclusion of EMSC parameters in the classification model has resulted in adequate accuracy (i.e., 80% or higher) for all species and a low level of systematic misclassification. It should be noted that, when establishing an FTIR classification tree, the

actual pollen taxonomy might not be optimal for the establishment of all nodes, because spectral similarities of pollen might differ in some cases from established phylogeny.²⁰

The presented model performance results are realistic and based on the best-case scenario in modeling when one can afford establishing calibration models and testing them on an independent test set. Cross-validated calibration results are more often reported in the literature on pollen classification.^{13,37–40} Unfortunately, these results are usually over-optimistic and do not give a real estimate on the model performance. Notwithstanding, cross-validation is a good option if only a limited number of spectra is available. In order to compare our results with the ones presented in the literature we report full (leave-one-out) CV results which we run on the full data set (390 spectra). When considering only chemical spectral data (data set A), the accuracy on the species level was 96.9%. When a sparse PLSR classification model was used, based on data set A at all levels except “order” level where data set C was used, the accuracy on the species level increased to 97.7%.

It is important to notice that there was no misclassification at subgenus level (Figure 4). It is apparent that μ FTIR enables classification of congeneric species as demonstrated by absence of misclassification between either two *Quercus* or two *Juniperus* species. Moreover, it is possible that the method could even enable classification at subspecies level. This has been tested by excluding one of the *Betula pendula* cultivar samples from the test set and using these 30 single grain spectra for an external validation. In that case, 77% of the kept away *B. pendula* cultivar grains were correctly identified as *B. pendula* cultivar, 17% were identified as *B. pendula* wild type, while 6% were not identified as *B. pendula*.

The plots of the regression coefficients for the Sparse PLSR model are displayed in Figure 5. It can be seen that, when considering only chemical spectral data (data set A), spectral differences between pollen orders (Poales, Fagales, and Pinales) are present in the whole spectral range (Figure 5A). The corresponding regression coefficients for a sparse PLSR classification model including EMSC parameters (data set C) are shown in Figure 5B. They are significantly less complex compared to the regression coefficients in Figure 5A, with only two main contributing bands that can be associated with lipids (C=O stretch at 1724 cm^{-1}) and proteins (Amide II band at 1537 cm^{-1}). Therefore, it is apparent that the significant reduction in misclassification at plant order level can be attributed to EMSC parameters. This improvement is expected because EMSC parameters are related to morphology of pollen, and pollen grains of the three plant orders have extremely different morphological features.²¹ As mentioned previously, a morphological approach at lower taxonomic levels (family and below) is limited due to lack of distinctive morphological features, and thus, inclusion of EMSC parameters may even result in worse classification. Therefore, inclusion of EMSC parameters (data set C) in the whole analysis (i.e., in all nodes of classification tree) has not improved overall accuracy. However, a sparse PLSR classification model built on either data set A or C at different nodes, has resulted in improved overall accuracy.

The regression coefficients at lower taxonomic levels are less complex than at order level and emphasize the advantage of Sparse PLS in enhancing the interpretation of the classification results. Discrimination between grass (*Bromus erectus*) and sedge (*Carex pendula*) is based on stronger lipid-related

absorption of sedge pollen at 1745 cm^{-1} (C=O stretch) and 1163 cm^{-1} (C–O stretch), and stronger protein-related absorption of grass pollen in the “protein” region (Figure 5C). Unfortunately, only two Poales species were included in this study, and thus, further studies with a larger pollen set are required. However, the result is in accordance with a previous study on 15 sedge (Cyperaceae) and 32 grass (Poaceae) pollens, which has shown that pollen of sedges have significantly increased lipid content when compared with the grasses.²⁰ Moreover, the previous study has shown that pollen of grass and sedge families have relatively low FTIR spectral variability compared to, for example cypress (Cupressaceae) and beech (Fagaceae) families.²⁰ Therefore, high accuracies for discrimination of grasses and sedges pollen are expected in general.

The regression coefficients for classification of Fagales pollen between birch (Betulaceae) and beech (Fagaceae) families show that the classification is based on “lipid” and “carbohydrate” regions. The regression coefficients for classification of Fagaceae pollen between oak (*Quercus*) and beech (*Fagus*) genera are even less complex and show that the principal difference is based on absorption in the “lipid” region at 1743 cm^{-1} (C=O stretch) (Figures 5E). Again, this is in accordance with previous spectroscopic studies on pollen composition.²⁰

Future Perspective of a μ FTIR-Based Pollen Classifier System. In 1996, Stillman and Flenley addressed the need for an automated pollen classifier system.⁴¹ The authors have compared the system with manual microscopy analysis and have established a set of goals for such system. Compared to existing technology and routines, an ideal automated system should allow (1) to count more pollen grains per sample, (2) to reduce analysis time per sample, (3) to reduce analyst bias and improve consistency, and (4) to determine pollen-type at higher levels, preferably at species level. A recently published review by Holt and Bennett on progress of automated pollen counters has extended these goals by adding that the system should require little modification of existing sample collection and preparation techniques, as well as impart no new limitations.⁷ Finally, the stated quantitative criteria specify that the automated system should be able to identify at least 40 different taxa with an accuracy of 95%. In comparison, it has been estimated that accuracy based on the morphological recognition by a human operator, used in aerobiological monitoring networks, is approximately 80%.¹³

Of all the published attempts on an automated palynology system, only seven attempts have included classification of 10 taxa or more.⁷ Most of the studies were performed on the morphology-based systems, by a combination of image processing and a learning-based classifier. The classification success rates of these systems, tested on a comparable range of taxa as in our study, were approximately 92%.^{37–40} These classification accuracies are all lower than 98% accuracy obtained in the here presented μ FTIR system. Unfortunately, all these studies were using cross-validation estimation of model performance, thus the reported accuracy is probably over-estimated. There has been only one study, by Kaya et al., using an independent set of data for validation of model performance, reporting an accuracy of 88%.⁴² This is lower the accuracy of 95% obtained here with μ FTIR, when comparable independent estimation of accuracy was performed. However, it is difficult to make comparison between the two studies since in the study by Kaya et al., plant taxa were not specified, although it can be

assumed that only pollen of entomophilous (insect-pollinated) plants was included. Because aeroallergenic pollen is predominantly anemophilous (wind-pollinated), with quite different chemical and morphological properties than entomophilous pollen,^{20,43} any further comparison of the results is irrelevant.

Up until now, there has been only one attempt with μ FTIR classification of pollen, classifying 11 taxa at an accuracy of 84%.¹³ The relatively low accuracy, compared to morphology-based systems, was due to relatively high spectral variability of pollen single grains within a sample. At that time, the spectral variability was attributed to chemical variability between the grains due to natural variation or due to pollen immaturity. However, a recent study has shown that the spectral variability is predominantly caused by Mie scattering artifacts.²³ This has been confirmed by the presented study, where reproducible and scatter-free spectra were obtained by conducting pollen measurements in embedding matrix. As expected, this has resulted in a significantly higher classification accuracy of 98%.

The presented embedding technique for μ FTIR measurements of pollen is adaptable to routine aeroallergen analyses. It is interoperable with the Hirst-type volumetric pollen trap, which is the most prevalent type of pollen sampler for air-sampling. In our experimental setup, the pollen samples were embedded in soft paraffin, which is a standard adhesive substance in Hirst-type pollen traps.⁴⁴ The polyethylene foil, which we used as substrate for the embedding matrix, is comparable to the polyester foil, a commonly used holder in pollen traps. The amount of paraffin used in μ FTIR measurements is comparable to the amount used in standard pollen monitoring measurements. Therefore, a future μ FTIR system for pollen classification can be in full compliance with the requirements. It is fully in agreement with the existing pollen samplers and preparation techniques, and it can provide identification of pollen up to species level. Moreover, identification and classification is bias-free and thus provides higher consistency than human experts. The μ FTIR methodology is devoid of labor-intensive sample preparation and data acquisition tasks, and the method does not require specific high-skilled expertise, as opposed to the traditional morphology-based pollen studies. In addition, μ FTIR has a high potential for further automation of measurement processes.

Both embedding materials, polyethylene and paraffin, are transparent in both the visible and the infrared spectral region. This is an extremely valuable property because it enables not only chemical characterization of pollen by μ FTIR microscopy but also morphological characterization by standard optical microscopy. It also implies that advancement in standard (optical microscopy-based) automated pollen classifier systems can be used in automation of the μ FTIR pollen classifier system. For example, as a first measurement step both systems require reliable robotics and image processing to locate pollen on an adhesive sampling tape. After locating pollen grains, a system could acquire an image and an infrared spectrum in successive order to obtain morphological and chemical information on pollen. Therefore, coupling of both approaches into an automated FTIR and optical-microscopies-based system should enable unprecedented classification accuracy.

CONCLUSIONS

The presented study clearly demonstrates that identification of single pollen grains can be based on transmittance FTIR microspectroscopy by using an embedding matrix. The classification analysis shows an enormous potential of the

μ FTIR method, surpassing state-of-the art approaches by a wide margin. Nevertheless, there is still ample space for further improvement of the method.

First, optimization of the FTIR instrumental setting should be focused on decreasing spectral acquisition time. In this study, we have employed strict instrumental settings with a large safe margin to avoid any possible information shortfall regarding spectral quality. This has resulted in a relatively long spectral acquisition time of approximately 1 min per grain. However, it is safe to assume that this measurement time can be halved, or even quartered, without significant information shortfall. It should be mentioned once more that the μ FTIR method saves significant time regarding sample preparation by avoiding any chemical pretreatment, as opposed to standard optical microscopy approach.

Second, embedding media may be further optimized taking into account both the infrared and the visible region of the electromagnetic spectrum. Although the optimization in the infrared is of importance to increase FTIR spectral quality, by decreasing signal-to-noise ratio, the visible region is of importance for aforementioned coupling of FTIR and optical microscopy.

Third, optimization of the classification model requires a 2-fold approach as well. Further measurements are required to evaluate and calibrate this method for a wide range of plant taxa. This includes assessment of pollen classification at subspecies level. Moreover, improvement of classification models require the buildup of a spectral library with a significantly wider range of species and populations than currently available. The results show that identification accuracy is taxon-specific, and thus, measurement of a wide range of species is a necessity for thorough assessment of methodology performance. Furthermore, for the classification of spectral data, we have used Sparse PLSR to establish robust calibration models.

The presented μ FTIR method for analysis of pollen is already outperforming the previously published studies of comparable scope. Further improvements are anticipated in the near future, resulting in a better taxonomic resolution, as well as faster and more economical measurements. The study shows enormous potential for pollen and in general bioparticle analysis, such as plant and fungal spores, dinoflagellate cysts and diatom microalgae. Moreover, the experimental setting allows implementation of μ FTIR spectroscopy in various fields of palynology, including aerobiology, palaeoecology, terrestrial and aquatic ecology, sedimentology, environmental archeology, forensics, and melissopalynology.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b03208.

The material includes details on sample measurement, spectral preprocessing, chemical and physical variability of pollen grains, and discrimination analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: boris.zimmermann@nmbu.no. Tel.: +47 6723 1576. Fax: +47 6496 5001.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research was supported by the European Commission through the Seventh Framework Programme (FP7-PEOPLE-2012-IEF project no. 328289).

■ REFERENCES

- (1) Cruz, A. A.; Popov, T.; Pawankar, R.; Annesi-Maesano, I.; Fokkens, W.; Kemp, J.; Ohta, K.; Price, D.; Bousquet, J. *Allergy* **2007**, *62*, 1–41.
- (2) Wallace, D. V.; Dykewicz, M. S.; Bernstein, D. I.; Bernstein, I. L.; Blessing-Moore, J.; Cox, L.; Khan, D. A.; Lang, D. M.; Nicklas, R. A.; Oppenheimer, J.; Portnoy, J. M.; Randolph, C. C.; Schuller, D.; Spector, S. L.; Tilles, S. A.; May, K. R.; Miller, T. A.; Druce, H. M.; Baroody, F. M.; Bernstein, J. A.; Craig, T. J.; Georgitis, J. W.; Pawankar, R.; Rachelefsky, G. S.; Settipane, R. A.; Skoner, D. P.; Stoloff, S. W. *J. Allergy Clin. Immunol.* **2008**, *122*, S1–S84.
- (3) Asher, M. I.; Montefort, S.; Bjorksten, B.; Lai, C. K. W.; Strachan, D. P.; Weiland, S. K.; Williams, H. *Lancet* **2006**, *368*, 733–743.
- (4) Besancenot, J. P.; Thibaudon, M. *Rev. Mal. Respir.* **2012**, *29*, 1238–1253.
- (5) Deak, A. J.; Makra, L.; Matyasovszky, I.; Csepe, Z.; Muladi, B. *Sci. Total Environ.* **2013**, *442*, 36–47.
- (6) Diaz, J.; Linares, C.; Tobias, A. *Aerobiologia* **2007**, *23*, 231–238.
- (7) Holt, K. A.; Bennett, K. D. *New Phytol.* **2014**, *203*, 735–742.
- (8) Holt, K.; Allen, G.; Hodgson, R.; Marsland, S.; Flenley, J. *Rev. Palaeobot Palyno* **2011**, *167*, 175–183.
- (9) Pappas, C. S.; Tarantilis, P. A.; Harizanis, P. C.; Polissiou, M. G. *Appl. Spectrosc.* **2003**, *57*, 23–27.
- (10) Fischer, G.; Braun, S.; Thissen, R.; Dott, W. *J. Microbiol. Methods* **2006**, *64*, 63–77.
- (11) Gottardini, E.; Rossi, S.; Cristofolini, F.; Benedetti, L. *Aerobiologia* **2007**, *23*, 211–219.
- (12) Watson, J. S.; Sephton, M. A.; Sephton, S. V.; Self, S.; Fraser, W. T.; Lomax, B. H.; Gilmour, I.; Wellman, C. H.; Beerling, D. J. *Photobio Sci.* **2007**, *6*, 689–694.
- (13) Dell'Anna, R.; Lazzeri, P.; Frisanco, M.; Monti, F.; Malvezzi Campeggi, F.; Gottardini, E.; Bersani, M. *Anal. Bioanal. Chem.* **2009**, *394*, 1443–1452.
- (14) Zimmermann, B. *Appl. Spectrosc.* **2010**, *64*, 1364–1373.
- (15) Fraser, W. T.; Scott, A. C.; Forbes, A. E. S.; Glasspool, I. J.; Plotnick, R. E.; Kenig, F.; Lomax, B. H. *New Phytol.* **2012**, *196*, 397–401.
- (16) Mularczyk-Oliwa, M.; Bombalska, A.; Kaliszewski, M.; Wlodarski, M.; Kopczynski, K.; Kwasny, M.; Szpakowska, M.; Trafny, E. A. *Spectrochim. Acta, Part A* **2012**, *97*, 246–254.
- (17) Pummer, B. G.; Bauer, H.; Bernardi, J.; Chazallon, B.; Facq, S.; Lendl, B.; Whitmore, K.; Grothe, H. *J. Raman Spectrosc.* **2013**, *44*, 1654–1658.
- (18) Fraser, W. T.; Watson, J. S.; Sephton, M. A.; Lomax, B. H.; Harrington, G.; Gosling, W. D.; Self, S. *Rev. Palaeobot Palyno* **2014**, *201*, 41–46.
- (19) Lahlali, R.; Jiang, Y. F.; Kumar, S.; Karunakaran, C.; Liu, X.; Borondics, F.; Hallin, E.; Bueckert, R. *Front. Plant Sci.* **2014**, *5*, DOI: 10.3389/fpls.2014.00747.
- (20) Zimmermann, B.; Kohler, A. *PLoS One* **2014**, *9*, e95417.
- (21) Zimmermann, B.; Tkalec, Z.; Mesic, A.; Kohler, A. *PLoS One* **2015**, *10*, e0124240.
- (22) Zimmermann, B.; Bagcioglu, M.; Sandt, C.; Kohler, A. *Planta* **2015**, *242*, 1237–1250.
- (23) Lukacs, R.; Blumel, R.; Zimmermann, B.; Bagcioglu, M.; Kohler, A. *Analyst* **2015**, *140*, 3273.
- (24) Bhargava, R. *Appl. Spectrosc.* **2012**, *66*, 1091–1120.
- (25) Kohler, A.; Kirschner, C.; Oust, A.; Martens, H. *Appl. Spectrosc.* **2005**, *59*, 707–716.
- (26) Zimmermann, B.; Kohler, A. *Appl. Spectrosc.* **2013**, *67*, 892–902.
- (27) Kohler, A.; Sule-Suso, J.; Sockalingum, G. D.; Tobin, M.; Bahrami, F.; Yang, Y.; Pijanka, J.; Dumas, P.; Cotte, M.; van Pittius, D. G.; Parkes, G.; Martens, H. *Appl. Spectrosc.* **2008**, *62*, 259–266.
- (28) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.
- (29) Le Cao, K. A.; Rossouw, D.; Robert-Granie, C.; Besse, P. *Stat. Appl. Genet. Mol.* **2008**, *7*, Article no. 35.
- (30) Karaman, I.; Qannari, E.; Martens, H.; Hedemann, M. S.; Knudsen, K. E. B.; Kohler, A. *Chemom. Intell. Lab. Syst.* **2013**, *122*, 65–77.
- (31) Bassan, P.; Byrne, H. J.; Lee, J.; Bonnier, F.; Clarke, C.; Dumas, P.; Gazi, E.; Brown, M. D.; Clarke, N. W.; Gardner, P. *Analyst* **2009**, *134*, 1171–1175.
- (32) Bassan, P.; Kohler, A.; Martens, H.; Lee, J.; Byrne, H. J.; Dumas, P.; Gazi, E.; Brown, M.; Clarke, N.; Gardner, P. *Analyst* **2010**, *135*, 268–277.
- (33) Bassan, P.; Kohler, A.; Martens, H.; Lee, J.; Jackson, E.; Lockyer, N.; Dumas, P.; Brown, M.; Clarke, N.; Gardner, P. *Journal of biophotonics* **2010**, *3*, 609–620.
- (34) Mohlenhoff, B.; Romeo, M.; Diem, M.; Wood, B. R. *Biophys. J.* **2005**, *88*, 3635–3640.
- (35) van Dijk, T.; Mayerich, D.; Carney, P. S.; Bhargava, R. *Appl. Spectrosc.* **2013**, *67*, 546–552.
- (36) Krimm, S.; Liang, C. Y.; Sutherland, G. B. B. M. *J. Chem. Phys.* **1956**, *25*, 549–562.
- (37) Ronneberger, O.; Schultz, E.; Burkhardt, H. *Aerobiologia* **2002**, *18*, 107–115.
- (38) Treloar, W. J.; Taylor, G. E.; Flenley, J. R. *J. Quaternary Sci.* **2004**, *19*, 745–754.
- (39) Dhawale, V. R.; Tidke, J. A.; Dudul, S. V. *Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Mysore, India, August 22–25, 2013; pp 79–84.
- (40) Ticay-Rivas, J. R.; del Pozo-Banos, M.; Travieso, C. M.; Arroyo-Hernandez, J.; Perez, S. T.; Alonso, J. B.; Mora-Mora, F. *Ifip Adv. Inf. Comm. Te* **2011**, *364*, 342–349.
- (41) Stillman, E. C.; Flenley, J. R. *Quat. Sci. Rev.* **1996**, *15*, 1–5.
- (42) Kaya, Y.; Erez, M. E.; Karabacak, O.; Kayci, L.; Fidan, M. *Grana* **2013**, *52*, 71–77.
- (43) Tanaka, N.; Uehara, K.; Murata, A. *J. Plant Res.* **2004**, *117*, 265–276.
- (44) Galan, C.; Smith, M.; Thibaudon, M.; Frenguelli, G.; Oteros, J.; Gehrig, R.; Berger, U.; Clot, B.; Brandao, R.; Grp, E. Q. W. *Aerobiologia* **2014**, *30*, 385–395.

Supporting Information

Analysis of allergenic pollen by FTIR microspectroscopy

B. Zimmerman*, V. Tafintseva, M. Bağcıoğlu, M. Høegh Berdahl and A. Kohler

Department of Mathematical Sciences and Technology, Faculty of Environmental Science and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway

*Corresponding author:

Boris Zimmermann

Department of Mathematical Sciences and Technology

Faculty of Environmental Science and Technology

Norwegian University of Life Sciences

Drøbakveien 31, 1432 Ås, Norway.

Tel: +47 6723 1576

Faks: +47 6496 5001

E-mail: boris.zimmermann@nmbu.no

Table of Contents	Page
Sample measurement	S-2
Spectral pre-processing: EMSC model	S-3
Chemical variability of pollen grains	S-4
Physical variability of pollen grains	S-5
Discrimination analysis	S-7
Supporting Information References	S-7

Sample measurement

The pollen samples were embedded in a soft paraffin between two sheets of 40 μm thick polyethylene foils, and placed in a bespoke sample holder. A sample ready for μFTIR measurement is shown in Fig. S-1. As can be seen in Figs. S-1C and S-1D, μFTIR setting enables measurement of individual pollen grain. For the presented study, only physically separated pollen grains were measured, with particular attention to avoid grains with depth overlap (z-axis stacking of grains). Therefore, each of the 360 recorded spectra belonged to a different individual grain.

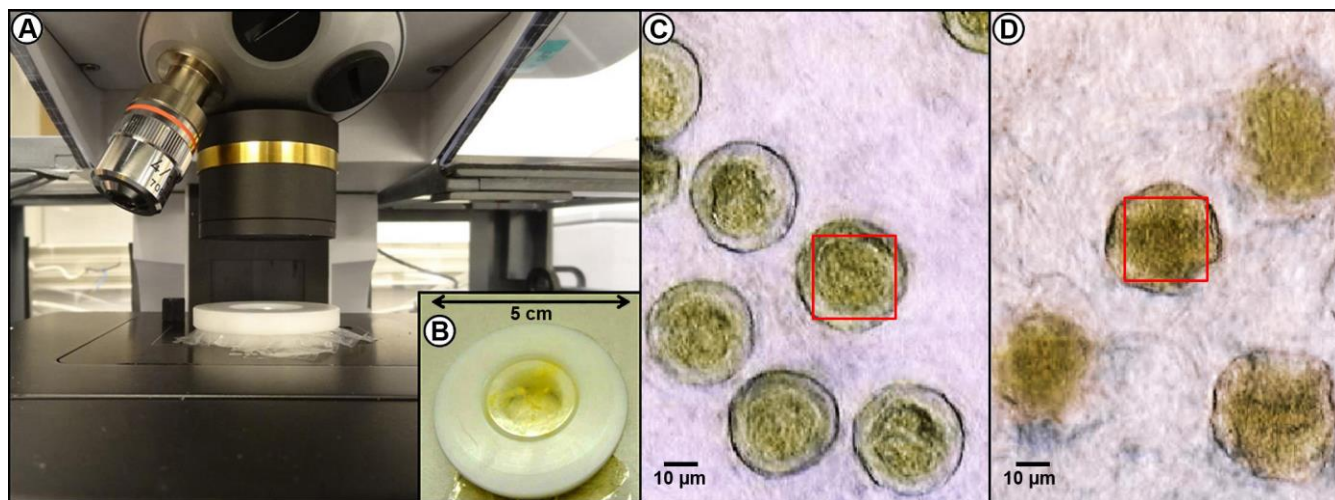


Fig. S-1. (A) Pollen sample on a moving bench and under objective of a HYPERION 3000 FTIR microscope, and (B) Sample holder with the prepared pollen sample in the polyethylene/paraffin embedding matrix. Optical microscope images taken during sample measurement with FTIR microscope: (C) *Cryptomeria japonica* grains, and (D) *Carpinus betulus* grains (out of focus grains are visible in lower left and upper right corners); 25 $\times$ 25 μm aperture of FTIR microscope is designated by the red square.

The measurement of pollen grains in the embedding medium is not as straightforward as the measurement of pollen grains on a microscope slide. Since bioparticles in the embedding medium are arranged within a three-dimensional matrix, it is improbable that they all have a common focal plane parallel with the foils (i.e. parallel with microscope stage). Usually the embedded bioparticles are arranged in different focal planes, and thus the focus of the FTIR microscope needs to be adjusted multiple times to measure all of the bioparticles. Fig. S-2 shows μFTIR spectra of pollen grains that were recorded using a single focal plane. Out of the ten presented spectra, half were out-of-focus. The out-of-focus spectra are easily distinguished by insignificant difference in absorbance values between the region where chemical absorbance of pollen is expected ('Amide I' region between 1700 and 1600 cm^{-1}) and the region where pollen grain does not have chemical absorbance ('Baseline' region between 2100 and 2000 cm^{-1}). Thus, the associated pollen grains were measured once more using a different focal plane in order to obtain high quality spectra. Quality of recorded spectra can be estimated by running the standard spectral

quality tests, as assessed in this study by calculating the signal-to-noise ratio of the Amide I signal as defined by OPUS/IR user manual.

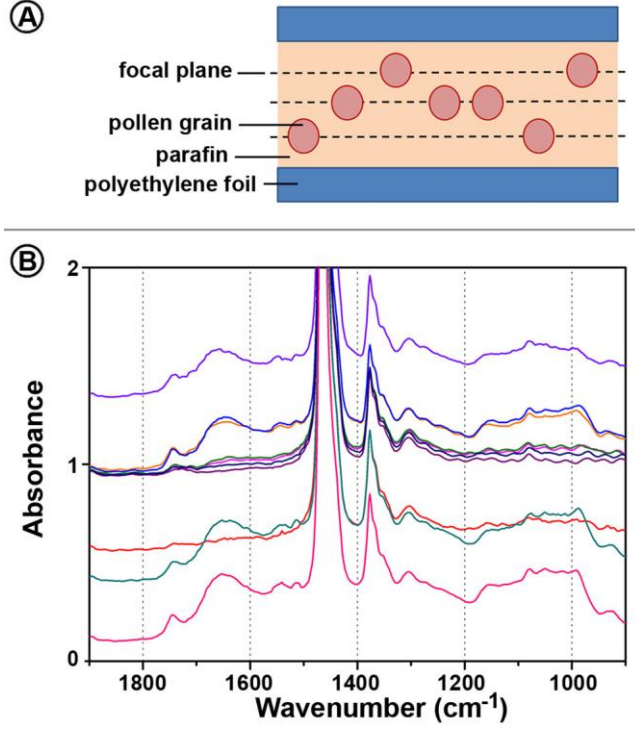


Fig. S-2. (A) Schematic drawing of μ FTIR measurements of pollen within paraffin matrix, showing arrangement of pollen grains in three different focal plains. (B) μ FTIR spectra obtained by measurement in paraffin and polyethylene embedding medium of ten individual *Betula pendula* pollen grains; half of the grains were in focus and half were out of focus during measurement

Spectral pre-processing: EMSC model

The EMSC model used in the pre-processing is defined by Eq. S-1, and the EMSC corrected spectra by Eq. S-2.

$$\mathbf{z}_i = b_i \mathbf{m} + a_i \mathbf{1} + d_i \tilde{\mathbf{v}} + e_i \tilde{\mathbf{v}}^2 + f_i \tilde{\mathbf{v}}^3 + \boldsymbol{\varepsilon}_i \quad (\text{S-1})$$

$$\mathbf{z}_{i,\text{Corr}} = (\mathbf{z}_i - a_i \mathbf{1} - d_i \tilde{\mathbf{v}} - e_i \tilde{\mathbf{v}}^2 - f_i \tilde{\mathbf{v}}^3) / b_i \quad (\text{S-2})$$

where $\mathbf{z}_i$ is a ‘measured’ spectrum (where $i = 1, \dots, N$; number of spectra in a data set N is 100), $\mathbf{z}_{i,\text{Corr}}$ is a corrected spectrum, b_i is a multiplicative parameter, $\mathbf{m}$ is a reference spectrum often presented by the mean spectrum for a given data set, $\mathbf{1} = [1, 1, 1, \dots, 1]$, a_i , d_i , e_i and f_i are constant, linear, quadratic and cubic parameters respectively, $\boldsymbol{\varepsilon}_i$ is a residual term, $\tilde{\mathbf{v}}$ is spectral range (wavenumbers).

Chemical variability of pollen grains

Pollen grains have unique vibrational spectra. Although pollen grains of a given type were collected from the same maternal plant, each grain within the sample is one genetically unique organism. Pollen development and its chemical composition involves contributions from both the sporophytic (maternal plant) and gametophytic (pollen) genomes^{1,2}. Therefore, each spectrum represents one biologically independent sample.

For principal component analysis (PCA), second derivative spectra were pre-processed by EMSC as described in the section about spectral pre-processing (dataset A). PCA analysis (Fig. S-3) shows substantial differences in the biochemical composition between pollen of angiosperms (Fagales and Poales) and conifers (Pinales). The previous studies have shown that the cypress (Cupressaceae) family, to which all four measured conifer taxa belong, has pollen with distinct spectral features^{3,4}. While the IR spectra of Cupressaceae pollen show dominant absorption due to vibrational bands in the ‘carbohydrate region’ (1200–900 cm^{-1}), pollen of angiosperms show strong IR absorption of amide I and amide II bands in the ‘protein region’ (1700–1520 cm^{-1}). Moreover, both pollen types differ in absorption of lipid related carbonyl stretch in the ‘lipid region’ (1750–1700 cm^{-1}), indicating differences in the chemical composition of pollen lipids. Cupressaceae pollens show strong absorption at approx. 1730 cm^{-1} , while angiosperms show strong absorption at approx. 1745 cm^{-1} .

Finally, PCA plots show well defined clustering at lower taxonomic levels, including species level. Of particular interest is a single clustering of all three *Betula pendula* samples. By using the first four principal components, spectra of samples belonging to distinct plant species can be separated quite well from each other. Therefore, μFTIR spectra of pollen show large potential for classification analysis.

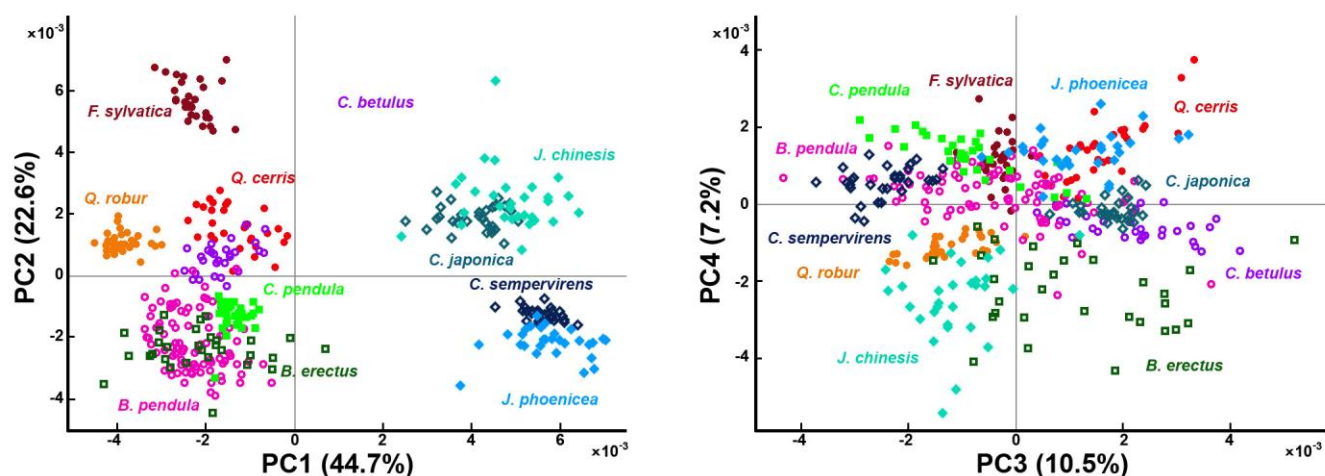


Fig. S-3. PCA score plots of a μFTIR spectral data set (30 spectra per sample; second derivative and EMSC corrected spectra – dataset A), with depiction of plant species. The percent explained variances for the first five PCs are 44.7, 22.6, 10.5, 7.2 and 3.8.

Physical variability of pollen grains

As stated above, a μ FTIR spectrum of a pollen grain includes, along with absorption signals of its chemical constituents, physical effects such as scattering distortions. These effects depend on sample morphology and the refractive index. In addition, variations in sample thickness introduce strong multiplicative effects. Pre-processing methods, such as derivative transformation of recorded data and extended multiplicative signal correction (EMSC), are often applied in order to reduce multiplicative and scatter effects, and to enhance chemical interpretation of the spectra.

EMSC is a model-based pre-processing method that allows explicit parameterization of physical information. It enables the separation of physical and chemical information from the spectral data ⁵. In order to consider physical information, EMSC parameters (dataset B) were obtained from the non-derived spectra.

Although EMSC estimates, containing baseline, background and other anomalous features, are often discarded in spectral analysis, they may contain information that is important for sample characterization ⁵. This is valid, in particular, for characterisation of pollen since physical information is closely related to a characteristic morphology of a bioparticle ⁴. Pollen morphology varies substantially in shape, size, and wall pattern across plant taxa. The chemical composition of pollen is complex, due to chemical variation of pollen grain substructures, such as grain wall layers (exine and intine), different type of cells (tube and generative cells) and intercellular material (cytoplasm and nuclei). This chemical composition of pollen varies substantially across plant taxa as well ³. Therefore, we expect that both chemical and physical information are of value for discrimination and classification of pollen.

Fig. S-4 shows the separation of physical and chemical information in μ FTIR spectra of pollen by an EMSC model with constant, linear, quadratic, and cubic parameters. The EMSC separates the recorded spectra (Fig. S-4A) into a “corrected chemical spectrum” including chemical absorption bands (Fig. S-4B), and EMSC estimates that contain physical information (Fig. S-4C). The EMSC scatter estimates include the complete “baseline” estimate (wavelike pattern in Fig. S-4C which is composed of the constant, linear and quadratic term of the EMSC model in Eq. S-2), and the effective optical path length (multiplicative parameter b in Fig. S-4C).

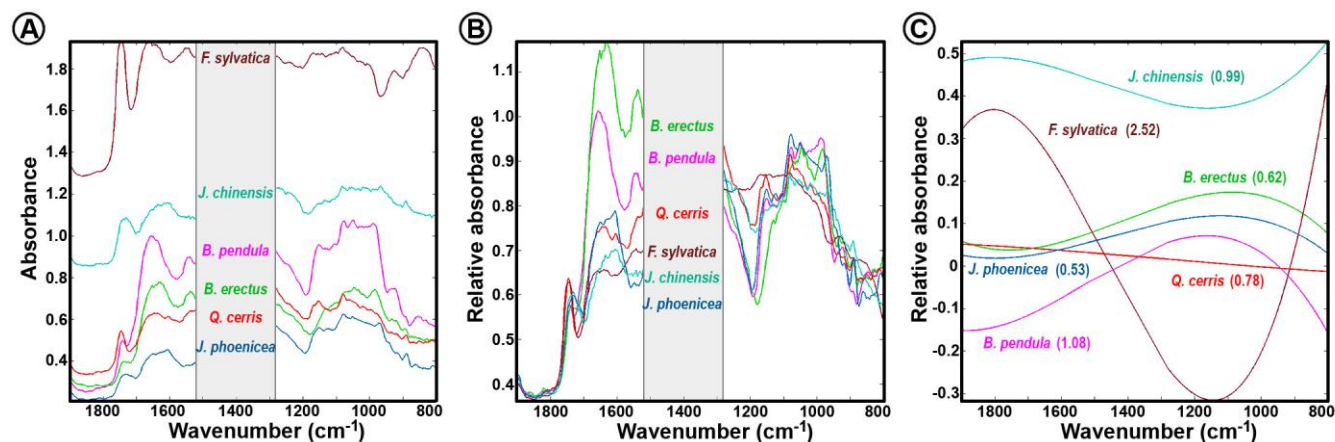


Fig. S-4. μ FTIR spectra of representative pollen grains of six different pollen species: (A) recorded spectra, (B) spectra after EMSC pre-processing, and (C) EMSC estimates (values of multiplicative parameter b are in parenthesis).

The graphical representation of EMSC parameters (dataset B) clearly shows that EMSC estimates are not completely random but to some extent specific for each taxon (Fig. S-5). Similar taxon-specific results for EMSC estimates were obtained previously on bulk measurement of spore and pollen samples by single reflectance attenuated total reflectance (SR-ATR) FTIR ⁴. It is expected, that a number of factors connected to pollen morphology of a measured pollen have impact on the resulting μ FTIR spectrum, even when chemical absorption is disregarded.

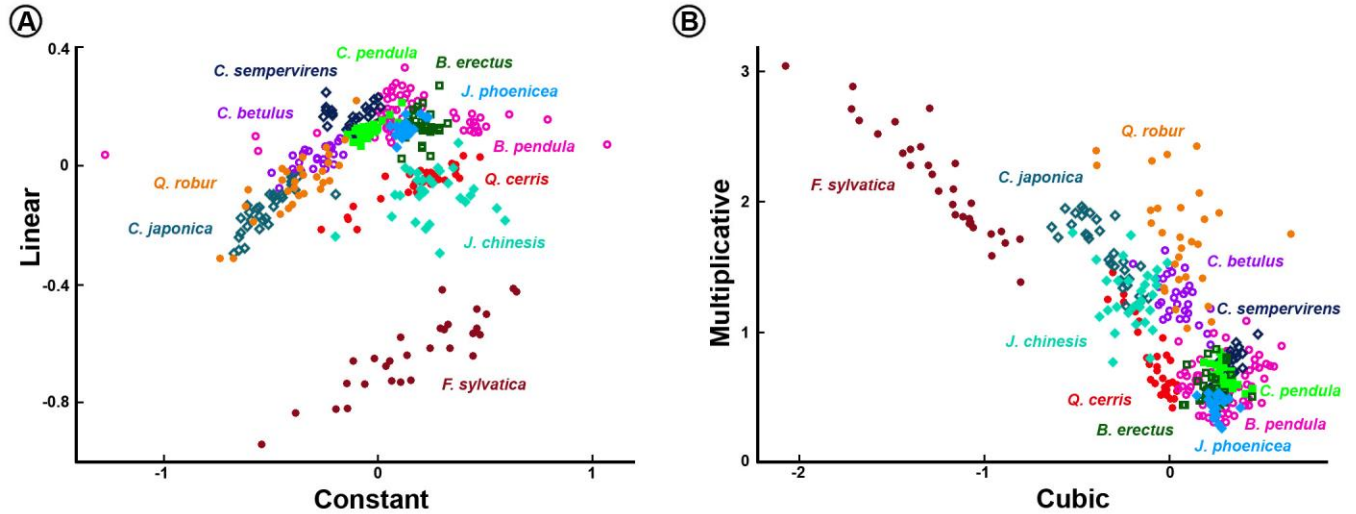


Fig. S-5. Plots of EMSC parameters (dataset B) of μ FTIR spectra (30 spectra per sample; EMSC corrected non-derived spectra), with depiction of plant species: (A) linear (d) and constant (a) parameters, and (B) multiplicative (b) and cubic (f) parameters.

One important factor is path length of infrared light through a pollen grain. Larger grains have longer optical path length, resulting in higher absorption values. The size of the measured grains varied from 15-25 μ m in diameter for Pinales, to 25-35 μ m for Fagales and Poales, with the exception of *Fagus sylvatica* that has significantly larger pollen grains with diameters of approx. 45 μ m. It can be seen that *F. sylvatica* pollen has completely different spectral scatter estimates compared to the rest of the samples (Fig. S-5). Although both *F. sylvatica* and *Quercus cerris* have rather similar chemical composition (Fig. S-3), characterised by large relative content of triglyceride lipids ³, their recorded spectra differ considerably regarding total absorbance (Fig. S-4). The high total absorbance for *F. sylvatica* grains is due to the relatively large size of *F. sylvatica* grains. The high absorbance values of *F. sylvatica* spectra indicate that the major part of the infrared light is completely absorbed by *F. sylvatica* grains.

One additional important factor is the variability of optical properties of pollen due to variation in chemical composition of pollen grain substructures. For instance, the two measured *Juniperus* species, *J. chinensis* and *J. phoenicea*, share similar morphology of pollen grains. However, their chemical composition is quite different, since *J. chinensis* has considerably larger relative content of triglyceride lipids than *J. phoenicea* ⁶. Triglyceride lipids are stored

predominantly within the pollen grain, in the form of intracellular spherical droplets, termed lipid bodies¹. Triglycerides exhibit strong absorption of infrared light, and therefore the difference in the relative content of triglyceride will lead to a significant difference related to multiplicative effect between the two *Juniper* species. Moreover, it can be seen that the two *Juniper* species show large differences in the spectral baseline (Fig. S-4). The baseline differences are most probably related to: 1) differences in optical densities due to differences in the constant part of the refractive index, and 2) differences in scattering properties due to Rayleigh scattering on lipid bodies and other organelles.

In general, EMSC estimates of pollen μ FTIR spectra have a range of taxon-specific values. As such, they offer an indirect physical measurement of a pollen sample that should be used for pollen characterization, alongside chemical absorption spectrum.

Discrimination analysis

Sensitivity and specificity of discrimination analysis for spectral data with included EMSC parameters (dataset C) is presented in Figure S-6.

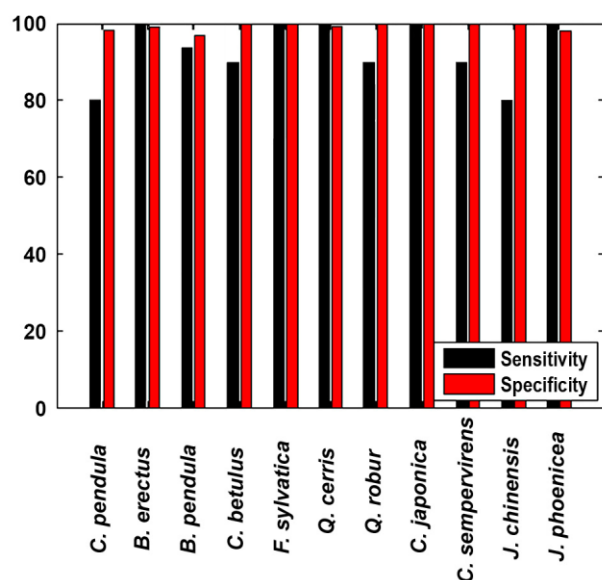


Fig. S-6. Identification result for Sparse PLSR analysis of pollen μ FTIR spectra, showing bar plots of sensitivity and specificity for each species (based on 10 spectra per sample). The result is based on second derivative spectral data with included EMSC parameters (dataset C).

Supporting Information References:

- (1) Piffanelli, P.; Ross, J. H. E.; Murphy, D. J. *Sex Plant Reprod* **1998**, *11*, 65-80.
- (2) Zhang, M.; Fan, J. L.; Taylor, D. C.; Ohlrogge, J. B. *Plant Cell* **2009**, *21*, 3885-3901.
- (3) Zimmermann, B.; Kohler, A. *Plos One* **2014**, *9*, e95417.
- (4) Zimmermann, B.; Tkalec, Z.; Mesic, A.; Kohler, A. *Plos One* **2015**, *10*, e0124240.
- (5) Kohler, A.; Kirschner, C.; Oust, A.; Martens, H. *Applied spectroscopy* **2005**, *59*, 707-716.
- (6) Zimmermann, B. *Applied spectroscopy* **2010**, *64*, 1364-1373.