Supplementary Materials for

A new non-destructive method to decipher the origin of organic matter in fossils using Raman Spectroscopy

Rossi *et al.*

*Corresponding author. Email: valentina.rossi@ucc.ie

This PDF file includes:

Supplementary Text
Figs. S1 to S13
References (90 to 97)

Other Supplementary Materials for this manuscript include the following:
Dataset S1 to S12

Supplementary Text

Principal Component Analysis: results of this study.

PCA of the entire experimental dataset (Fig. S11) identifies four broad groups: 1) kerogen-like Raman signatures (i.e., samples exhibiting D and G bands: melanins and experimentally matured samples); 2) carotenoid-rich samples; 3) chitinous samples and 4) keratinous samples. Note that in the PCA chemospace, chitinous and keratinous samples plot separately but closely, whereas they are better separated in the LDA chemospace (Fig. 2). The major discriminants that explain the variation in PC1 are the wavenumbers that form the D and G bands (1330 – 1355 cm^{-1} and 1570 – 1600 cm^{-1}, respectively, the saddle (ca. 1355 – 1500 cm^{-1}; refer to Fig. S1 for Raman nomenclature) and peaks from 900 – 1010 cm^{-1} (corresponding to carotenoids). The major discriminants that explain variation in PC2 are peaks assigned to C–N bonds (ca. 950 cm^{-1}) in chitin, phenylalanine (ca. 1000 cm^{-1}) and amide I (ca. 1620 cm^{-1}) in keratin and C=C bonds (ca. 1155 cm^{-1}) in carotenoids (Fig. S11). This chemospace does not discriminate samples in Group 1 (i.e., untreated melanin-rich and all matured samples).

Subsequent PCA of only those samples in Group 1 also fails to separate untreated melanin-rich and (all) experimental samples (Fig. S12). The loadings show differences in the region of the D band (1330 – 1355 cm^{-1}) and saddle (1355 – 1500 cm^{-1}) in PC1 and in the region of the G band (1570 – 1600 cm^{-1}) in PC2, but these differences are not captured in the chemospace. Note that the loadings plot (Fig. S12) is noisier than the example in Fig. S12 (PCA on entire dataset). The extensive overlap among the groups demonstrates a high degree of similarity among the Raman signatures of the Group 1 samples.

The plot of all melanin-rich untreated, experimental and fossil samples (PC1 = 51.5%, PC2 = 21.7, PC3 = 12.9%; Fig. S13). In this PCA scatterplot, only matured cyanobacterial film, matured leaf and fossilized plant samples occupy distinct regions of chemospace; all other samples overlap extensively. The loadings plot reveals that the major components that define the chemospace are the wavenumbers that form the D band (1330 – 1355 cm^{-1}) and the saddle (1355 – 1500 cm^{-1}) (collectively corresponding to negative loadings in PC1) and the G band (corresponding to negative loadings in PC2). The results of the PCA show that this approach to the analysis of our Raman data prevents identification of significant differences among the samples.

Our study demonstrates that for our dataset, LDA (but not PCA) of data from the entire spectrum successfully discriminates most known groups (see Fig. 2A and 5A). For those groups that were not discriminated using the above approach, LDA analysis of calculated Raman parameters yields scatterplots with clearly separated groups (see Discussion in main text).

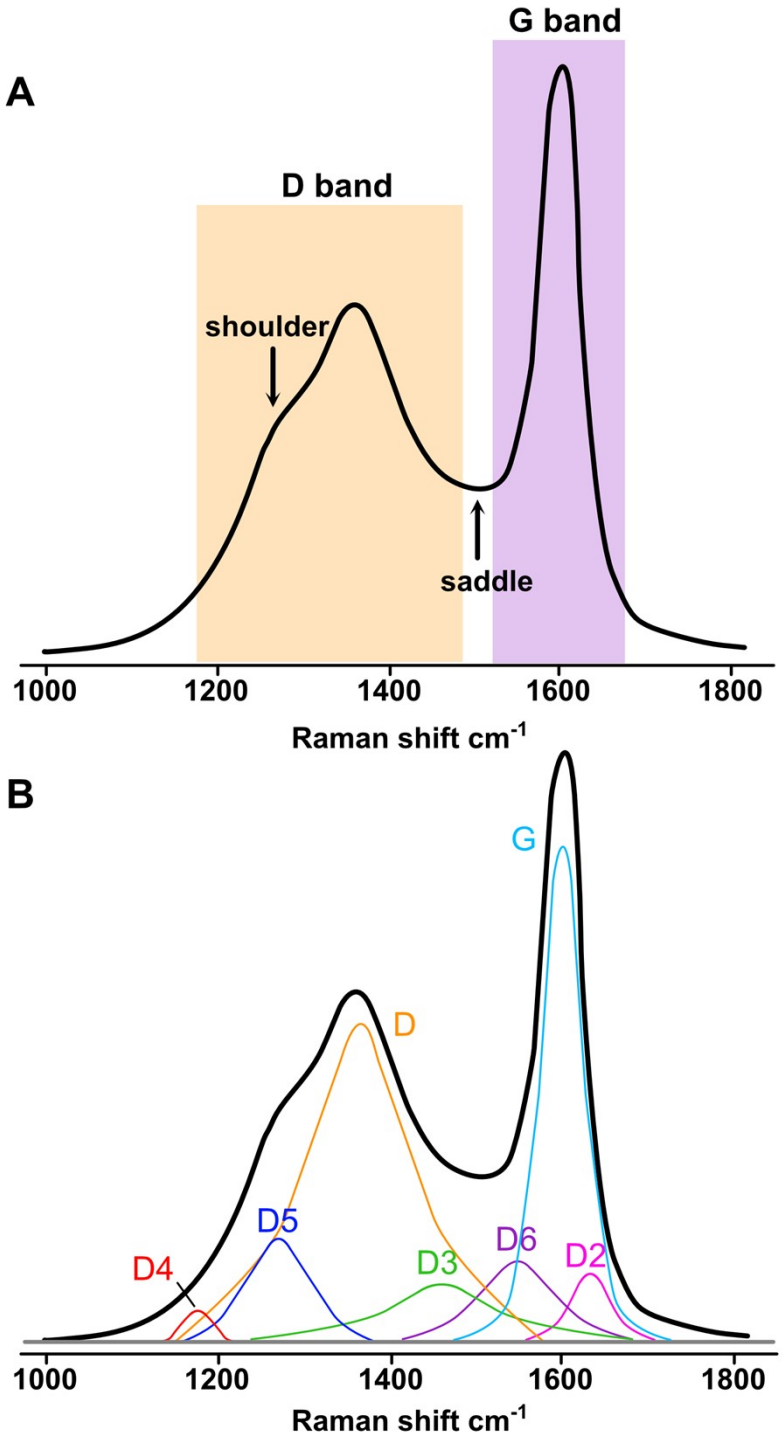

Fig. S1 – Anatomy and nomenclature of Raman spectra for kerogen (modified from Schito et al., 2023). (**A**) Typical Raman spectrum of low-grade organic matter and typical spectral features. (**B**) Peak deconvolution of the D- and G bands showing diagnostic secondary peaks.

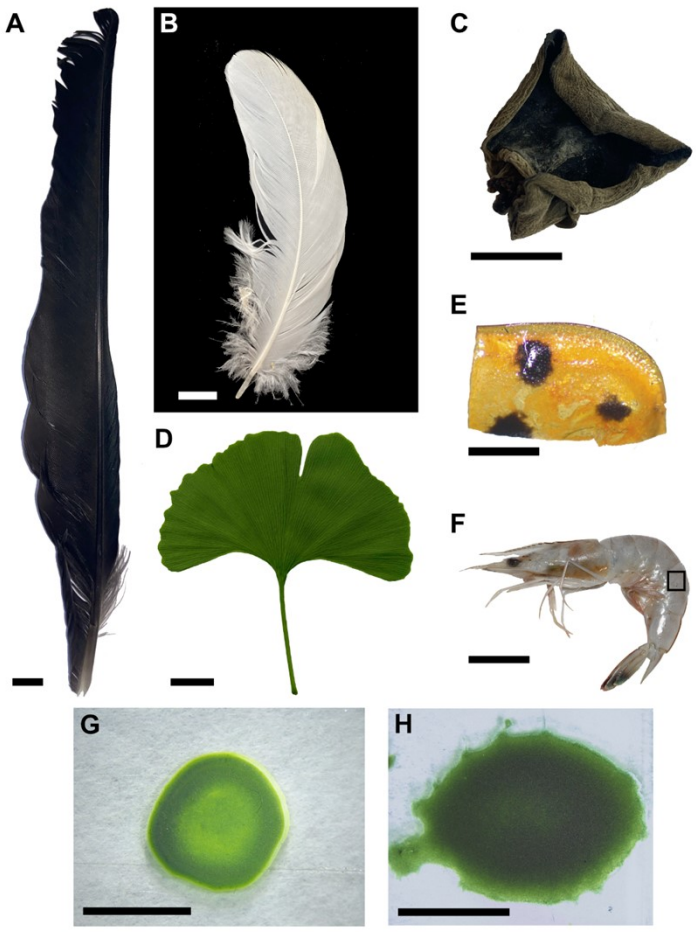

Fig. S2 – Extant samples used for thermal maturation experiments. (**A**) black feather (rook, *Corvus frugilegus*). (**B**) white feather (little egret, *Egretta garzetta*). (**C**) melanotic fungus (Judas's ear; *Auricularia auricola*). (**D**) leaf (*Gingko biloba*). (**E**) insect cuticle (elytron; ladybird, Coccinellidae). (**F**) shrimp cuticle (pink shrimp; *Pandalus borealis*). (**G**) green alga (*Chlorella sp.*). (**H**) cyanobacteria (*Nostoc punctiformes*). Scale bars: 10 mm (**A–D**, **F**); 1 mm (**E**); 5 mm (**G, H**).

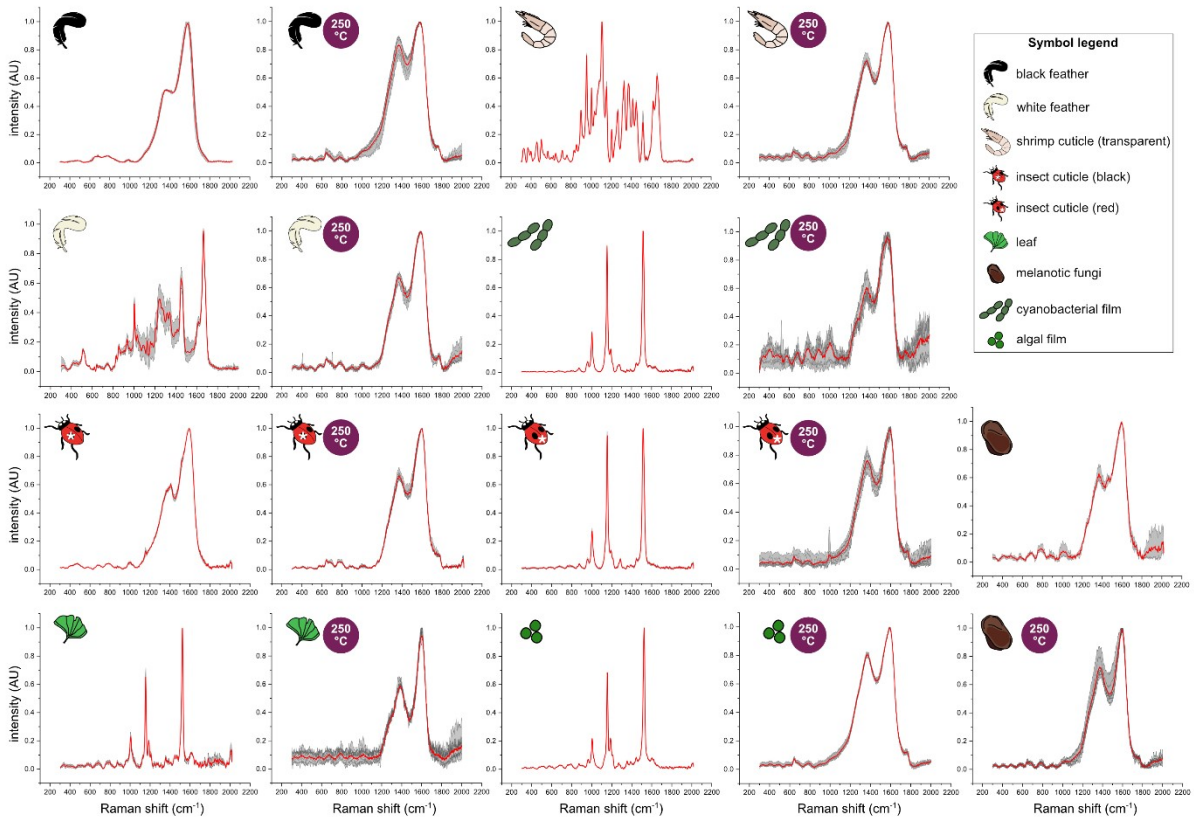

Fig. S3 – Raman spectra for extant and experimentally matured samples. Red: averaged spectra; dark grey: replicate spectra (n = 9); grey shaded area represents minimum and maximum intensity values.

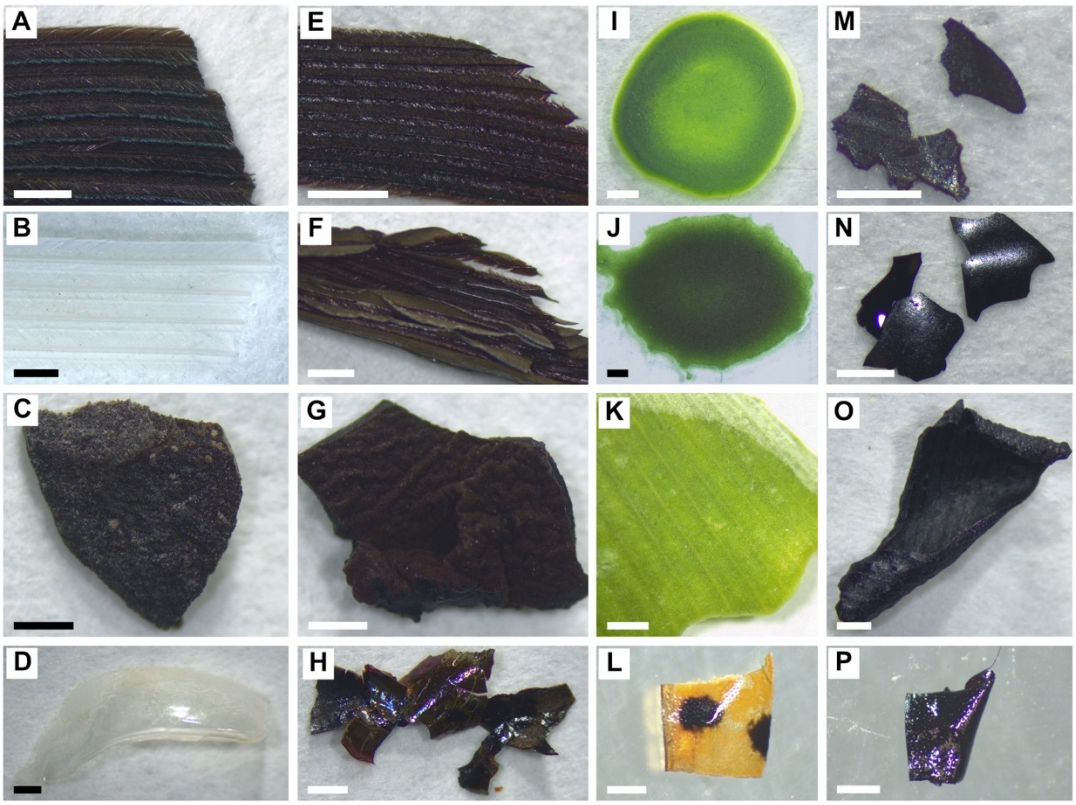

Fig. S4 – Photographs of extant tissue samples pre- and post-maturation (250°C). (**A – D**, **I – L**) Untreated samples. (**E – H**, **M – P**) Experimentally matured samples. (**A, E**) black feather (rook, *Corvus frugilegus*). (**B, F**) white feather (little egret, *Egretta garzetta*). (**C, G**) melanotic fungus (Judas's ear; *Auricularia auricola*). (**D, H**) shrimp cuticle (pink shrimp; *Pandalus borealis*). (**I, M**) green algal film (*Chlorella sp.*). (**J, N**) cyanobacterial film (*Nostoc punctiformes*). (**K, O**) leaf (*Gingko biloba*). (**L, P**) insect cuticle (elytron; ladybird, Coccinellidae). Scale bars: 1 mm (**A – K**, **N, O**); 0.5 mm (**L, M, P**).

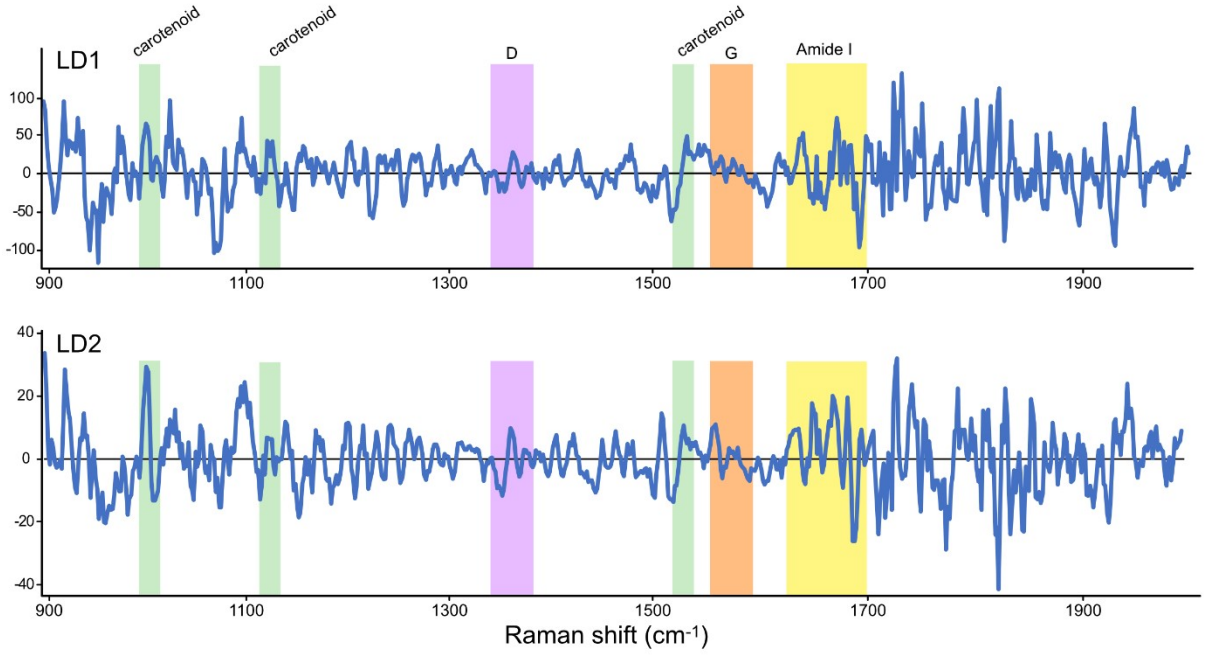

Fig. S5 – Loadings associated with the Linear Discriminant Analysis (LDA) chemospace plot in Fig 2A. Positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

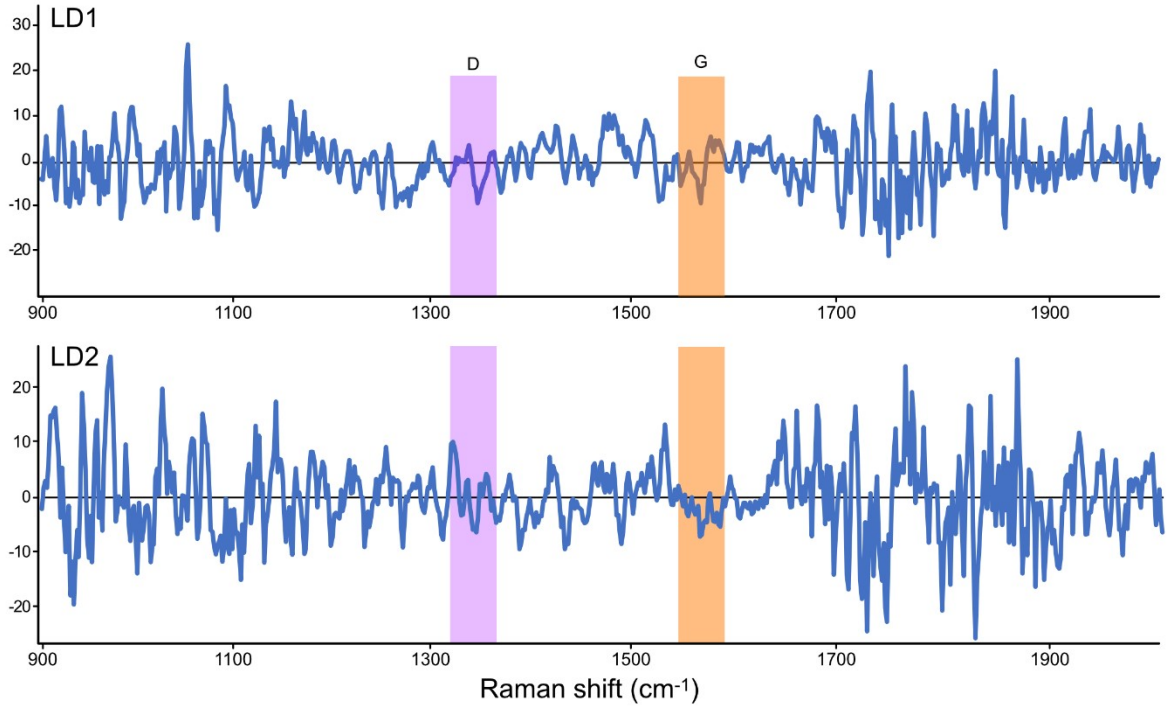

Fig. S6 – Loadings associated with the LDA chemospace plot in Fig 2B. Positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

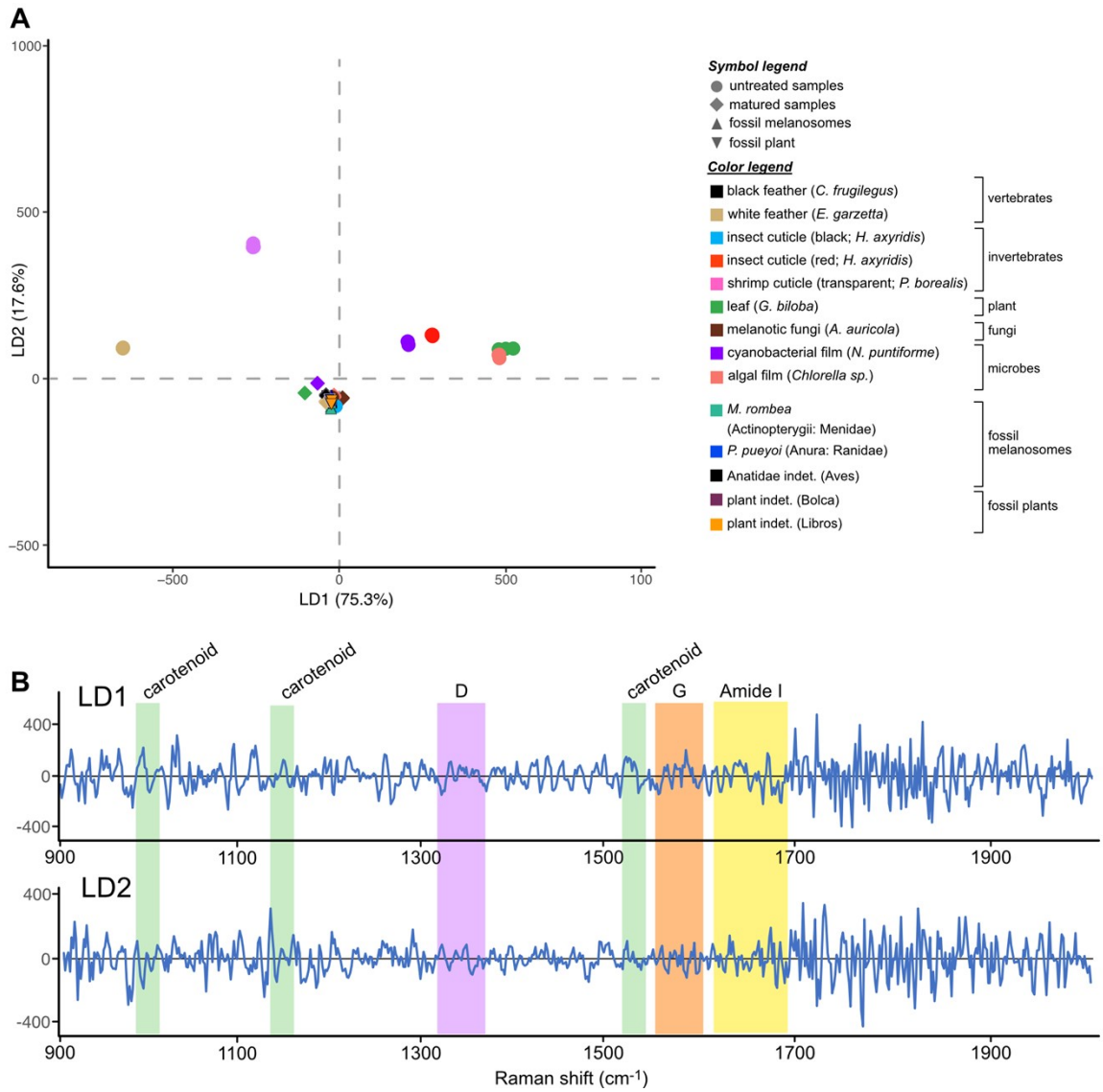

Fig. S7 – Linear Discriminant Analysis chemospace plot of Raman spectral data for all extant, experimentally matured and fossil samples. (**A**) chemospace; (**B**) loadings. Positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

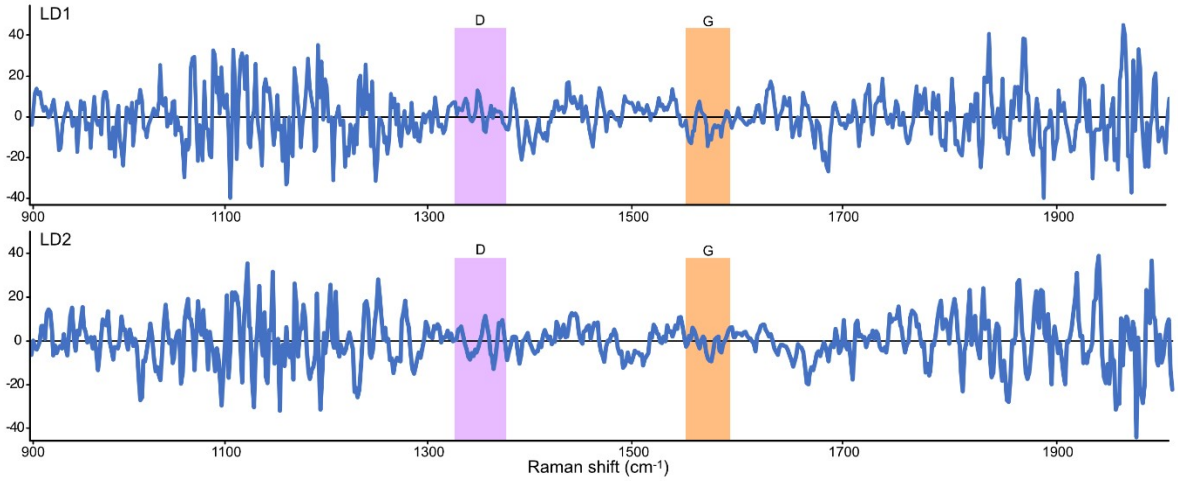

Fig. S8 – Loadings associated with the LDA chemospace plot in Fig 4A. Positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

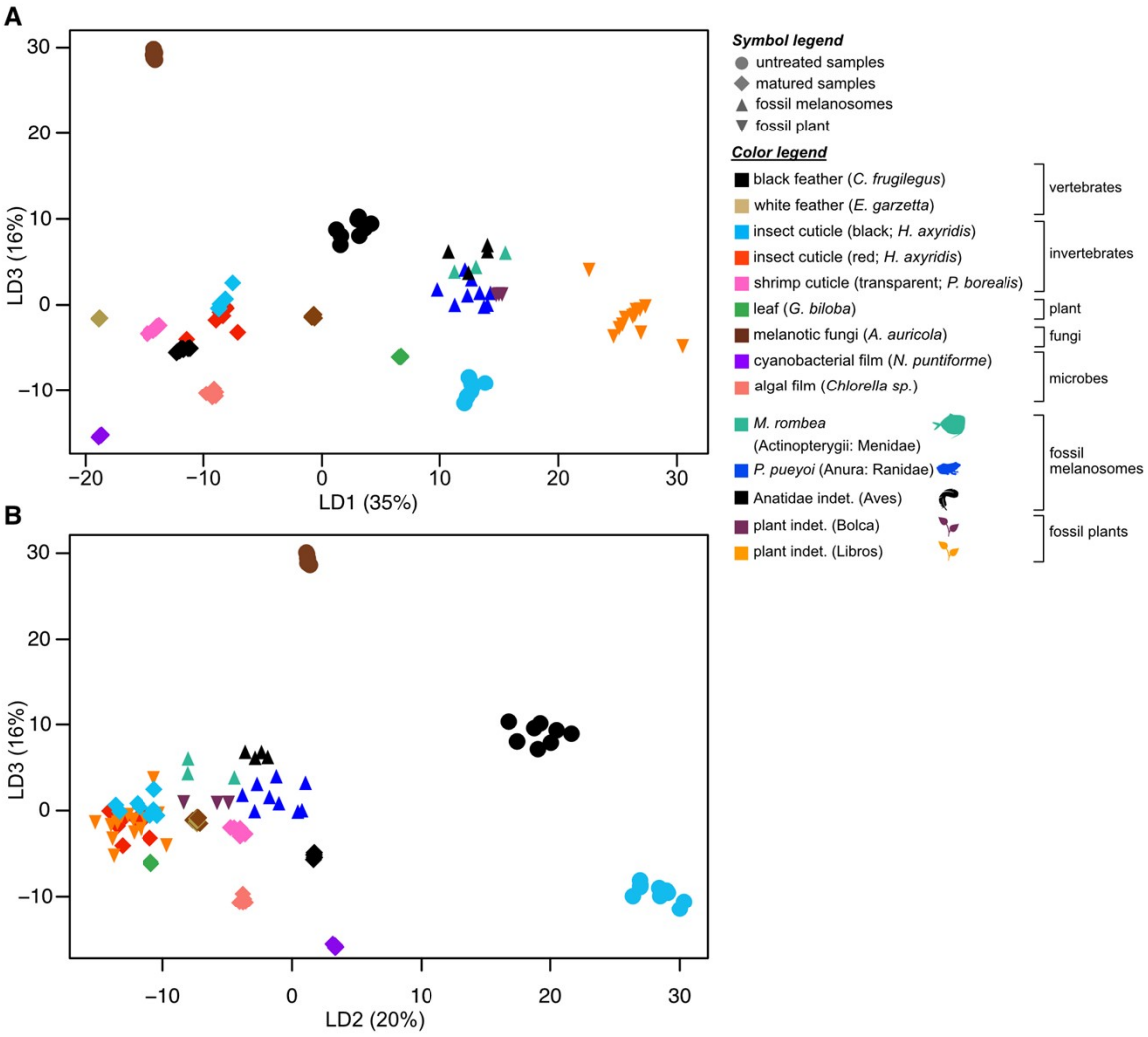

Fig. S9 – Supplementary Linear Discriminant Analysis scatterplots. These plots show the distribution of groups in chemospace for LD1 versus LD3 (**A**) and LD2 versus LD3 (**B**). The chemospace for LD1 versus LD2 is shown in Fig. 4**A**.

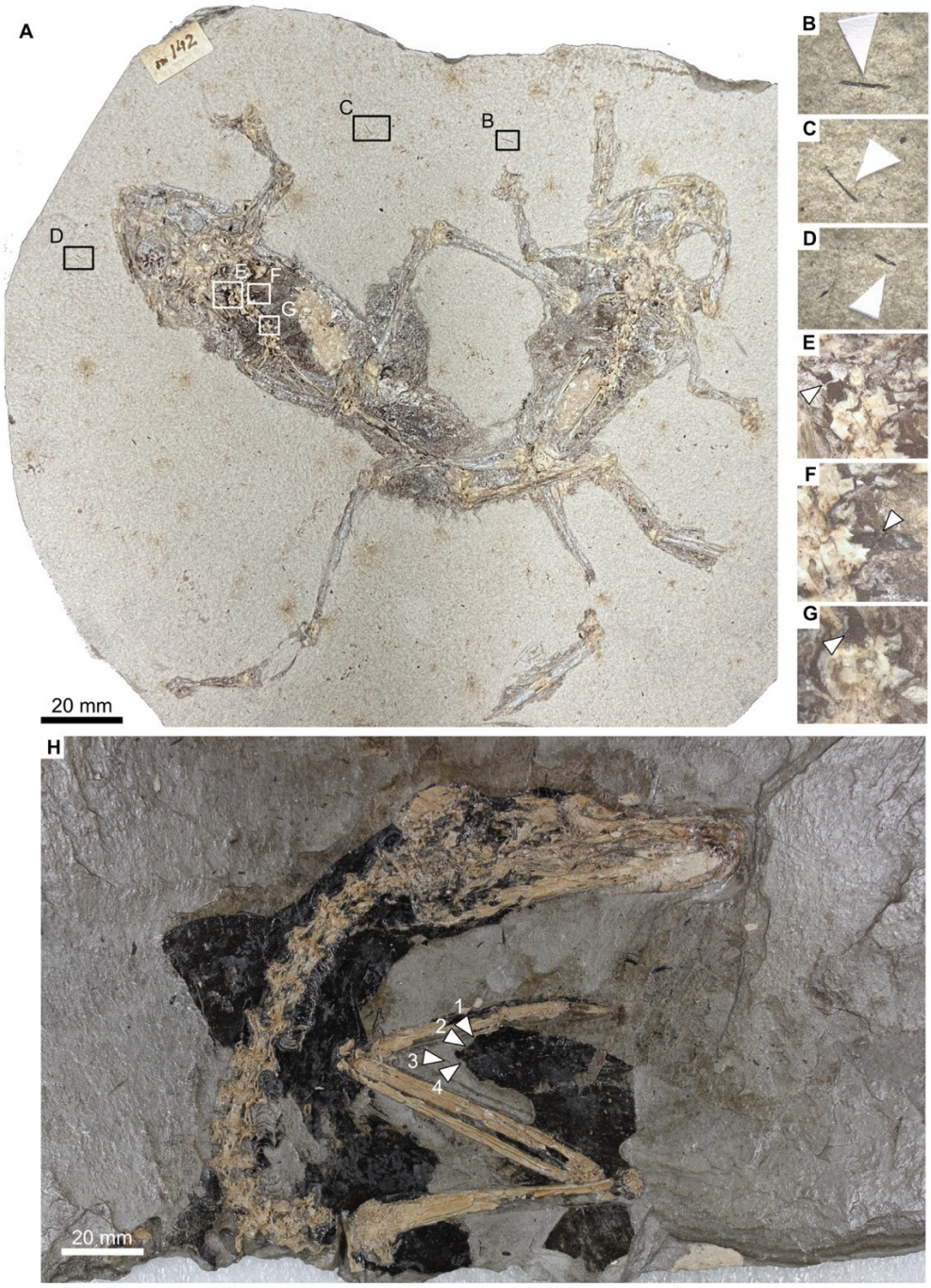

Fig. S10 – Fossil specimens from the Libros Konservat-Lagerstätte. (**A**) *Pelophylax pueyoi* (CKGM-m142; Anura). The specimen on the left was sampled. (**B – D**) indeterminate plant fragments analyzed in this study. (**E – G**) detail of the sample location of abdominal melanosome-rich tissue. (**H**) Anatidae indet. (MNCN 2.017 AV-001-a; Aves). Samples numbered 1 and 2 are melanosome-rich feather tissues; samples 3 and 4 are from indeterminate plant fragments.

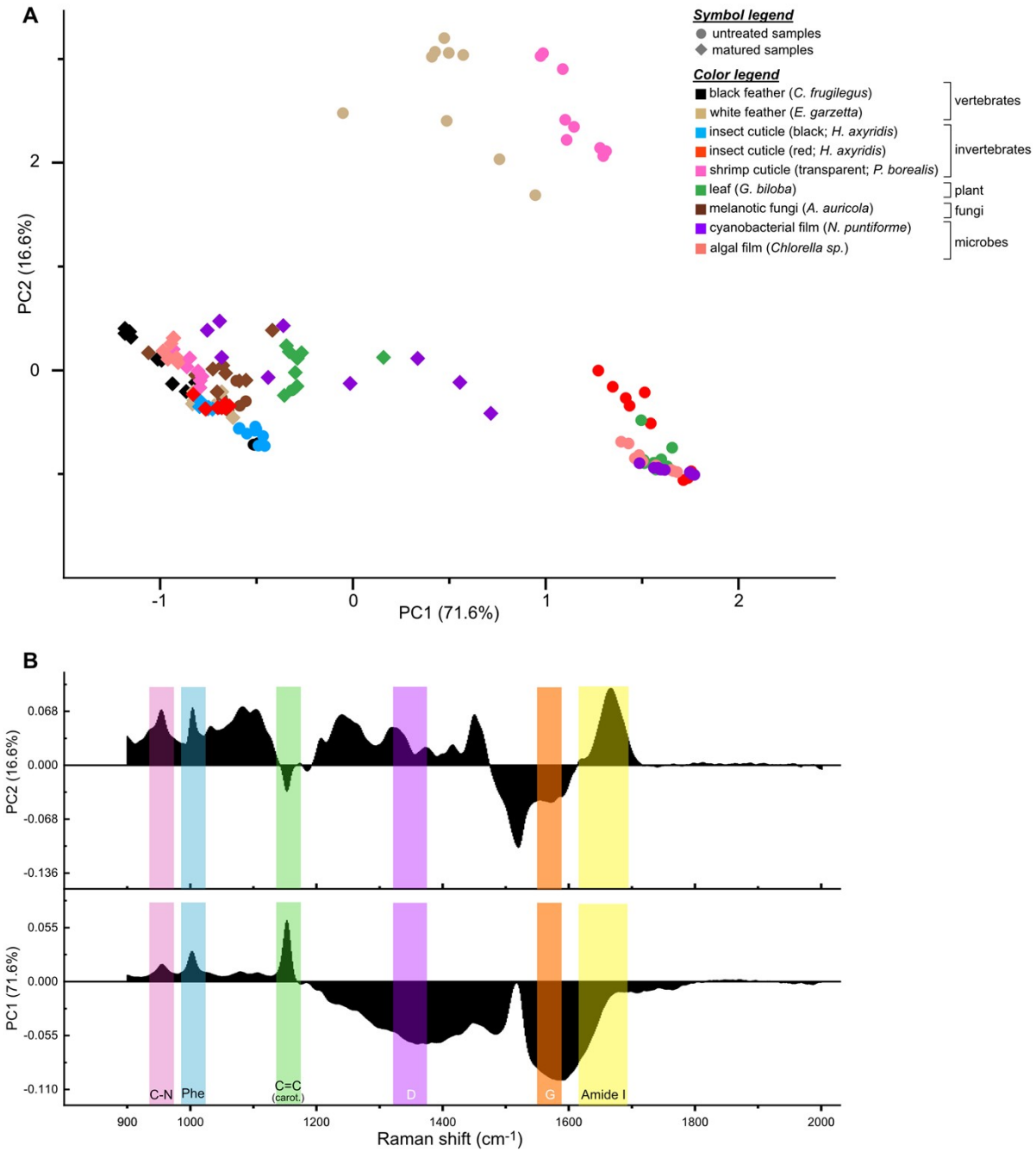

Fig. S11 – Principal Component Analysis (PCA) of Raman data for all untreated and experimentally matured samples. (**A**) Chemospace plot showing good separation among samples rich in keratin, chitin and carotenoids. Melanin-rich samples and experimentally matured samples show extensive overlap. (**B**) Principal Component loadings; positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details). Colour bands denote the position of key Raman bands; phenylalanine (Phe), carotenoid (carot.).

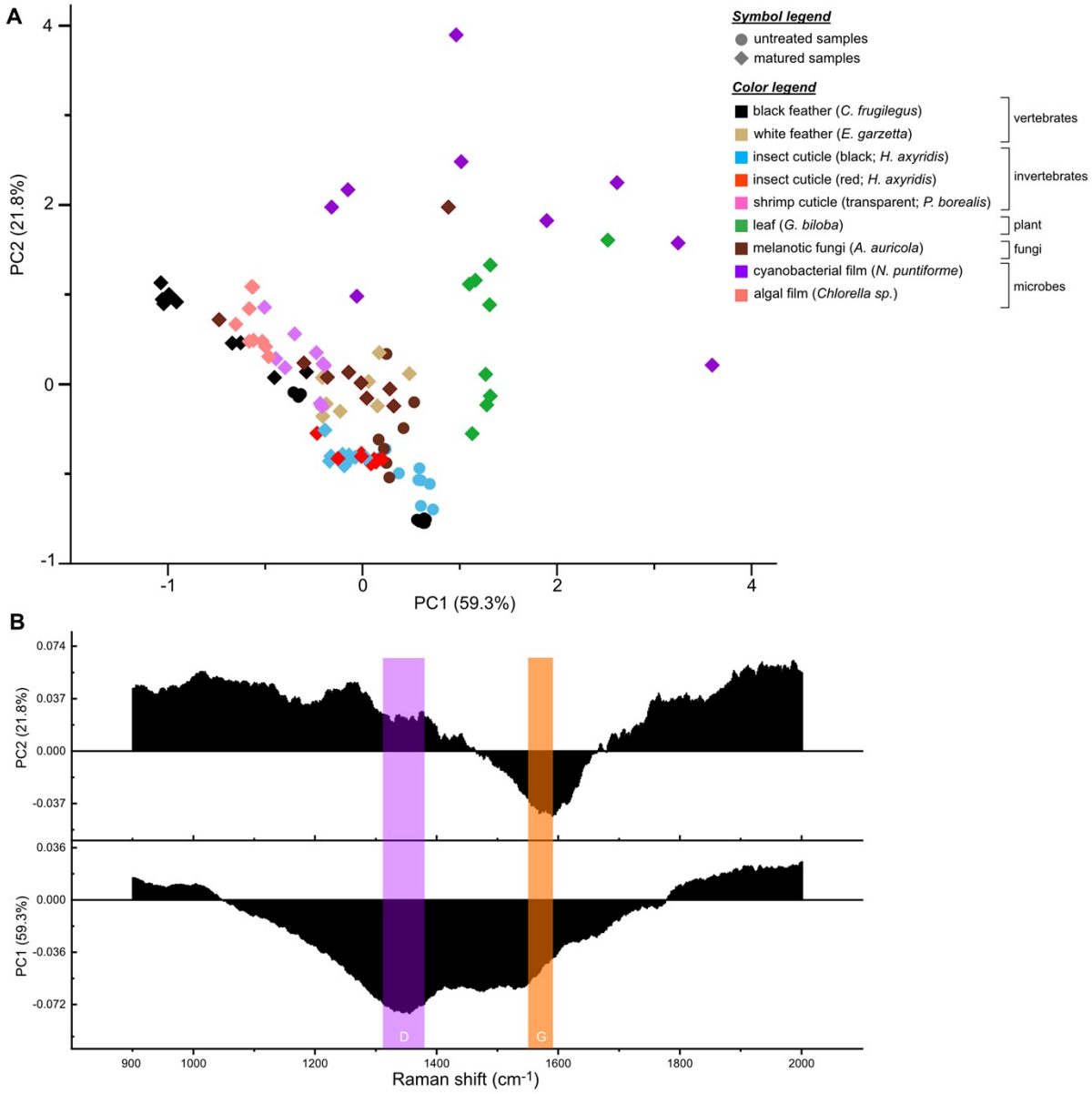

Fig. S12 – Principal Component Analysis (PCA) of Raman spectral data for untreated melanin-rich samples and matured samples. (**A**) Melanin-rich and experimentally matured samples overlap extensively in chemospace. (**B**) PC loadings; positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

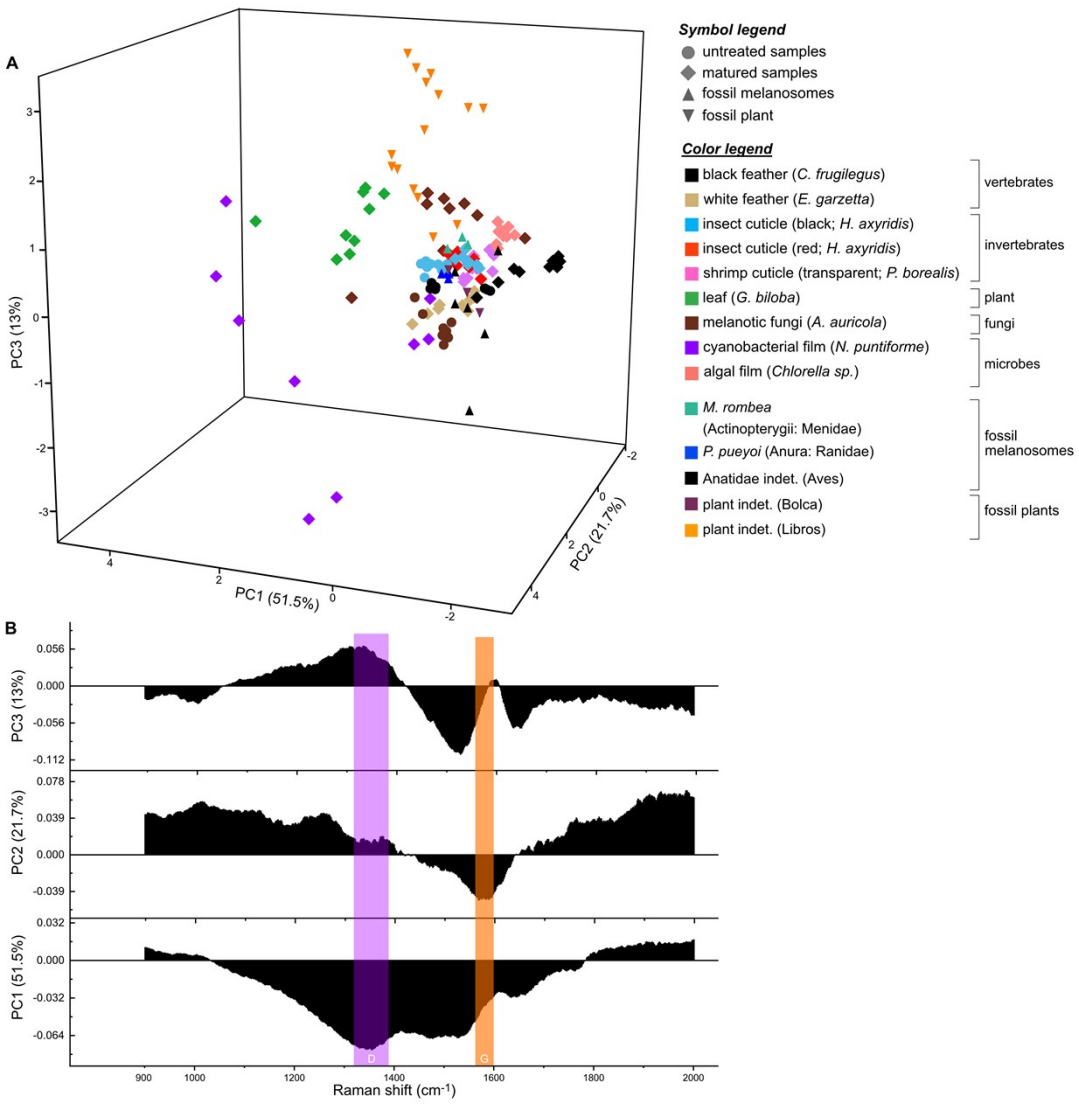

Fig. S13 – Principal Component Analysis (PCA) of raw Raman data for all untreated melanin-rich, matured and fossil samples. (**A**) Melanin-rich, experimentally matured and fossil samples overlap extensively in the chemospace. (**B**) PC loadings; positive and negative values in the line plots identify the variables (i.e., wavenumbers) responsible for the separation among groups in the chemospace (see supplementary text for details).

Supplementary Datasets

Dataset S1 - List of major Raman bands and peaks in the samples analyzed. (Phe) phenylalanine.

Dataset S2 - Raman spectra (baseline corrected and normalized) for all samples in the dataset.

Dataset S3 - LDA loadings for the chemospace shown in Fig. 2A. LD1 and LD2 loadings are represented as line plots in Fig. S6.

Dataset S4 - LDA loadings for the chemospace shown in Fig. 2B. LD1 and LD2 loadings are represented as line plots in Fig. S7.

Dataset S5 - Raman parameters extracted from the deconvolution of the D and G bands. D- and G frequency (i.e., the wavenumber defining the center of a peak), D-FWHM (full width at half maximum height), G-FWHM, R1(ID/IG; I denotes intensity of the peak), Aratio (areaD/areaG), RBS (Raman Band Separation: Dfrequency-Gfrequency), wD/wG (where w denotes FWHM), D2-FWHM (D2-full width at half maximum height), D4-FWHM (D4-full width at half maximum height) and D6-FWHM (D6-full width at half maximum height).

Dataset S6 - LDA loadings for the chemospace shown in Fig. 3. D- and G frequency (i.e., the wavenumber defining the center of a peak), D-FWHM (full width at half maximum height), G-FWHM, R1(ID/IG; I denotes intensity of the peak), Aratio (areaD/areaG), RBS (Raman Band Separation: Dfrequency-Gfrequency), wD/wG (where w denotes FWHM), D2-FWHM (D2-full width at half maximum height), D4-FWHM (D4-full width at half maximum height) and D6-FWHM (D6-full width at half maximum height).

Dataset S7 - Summary of the results of the ANOVA-type statistical analyses. The Kruskal-Wallis test was used for datasets were the assumption of normality was not met. The Welch ANOVA test was used when the assumption of normality and equality of variance were not met. Bold font denotes p-values < 0.05 (i.e., statistically significant).

Dataset S8 - Loadings for the LDA chemospace of untreated, matured and fossil samples shown in Fig. S10.

Dataset S9 - Loadings for the LDA chemospace including untreated melanin-rich, experimentally matured and fossil samples; chemospace shown in Fig. 4A.

Dataset S10 - Raman parameters extracted from the deconvolution of the D and G bands in spectra for fossil samples. D- and G frequency (i.e., the wavenumber defining the center of a peak), D-FWHM (full width at half maximum height), G-FWHM, R1(ID/IG; I denotes intensity of the peak), Aratio (areaD/areaG), RBS (Raman Band Separation: Dfrequency-Gfrequency), wD/wG (where w denotes FWHM), D2-FWHM (D2-full width at half maximum height), D4-FWHM (D4-full width at half maximum height) and D6-FWHM (D6-full width at half maximum height).

Dataset S11 - LDA loadings for the chemospace shown in Fig. 3. D- and G frequency (i.e., the wavenumber defining the center of a peak), D-FWHM (full width at half maximum height), G-FWHM, R1(ID/IG; I denotes intensity of the peak), Aratio (areaD/areaG), RBS (Raman Band Separation: Dfrequency-Gfrequency), wD/wG (where w denotes FWHM), D2-FWHM (D2-full width at half maximum height), D4-FWHM (D4-full width at half maximum height) and D6-FWHM (D6-full width at half maximum height).

Dataset S12 - List of secondary peaks derived from the deconvolution of the D and G bands in spectra for matured and fossil samples. FWHM=Full Width Half Maximum.

Supplementary References

90. A. Ditta, H. Nawaz, T. Mahmood, M. Majeed, M. Tahir, N. Rashid, M. Muddassar, A. Al-Saadi, H. Byrne, Principal components analysis of Raman spectral data for screening of Hepatitis C infection. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **221**, 117173 (2019).
91. S. Guo, P. Rösch, J. Popp, T. Bocklitz, Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications. *Journal of Chemometrics* **34**, e3202 (2020).
92. X. He, Y. Liu, S. Huang, Y. Liu, X. Pu, T. Xu, Raman spectroscopy coupled with principal component analysis to quantitatively analyze four crystallographic phases of explosive CL-20. *RSC Advances* **8**, 23348–23352 (2018).
93. P. Buzzini, J. Curran, C. Polston, Comparison between visual assessments and different variants of linear discriminant analysis to the classification of Raman patterns of inkjet printer inks. *Forensic Chemistry* **24**, 100336 (2021).
94. H. Li, Y. Ren, F. Yu, D. Song, L. Zhu, S. Yu, S. Jiang, S. Wang, Raman microspectral study and classification of the pathological evolution of breast cancer using both principal component analysis-linear discriminant analysis and principal component analysis-support vector machine. *Journal of Spectroscopy* **2021**, 1–11 (2021).
95. M. Lasalvia, V. Capozzi, G. Perna, A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra. *Applied Sciences* **12**, 5345 (2022).
96. I. T. Jolliffe, "Principal component analysis for special types of data" in Principal Component Analysis (Springer, New York, ed. 2, 2002), pp 338–372.
97. N. Gerhardt, M. Birkenmeier, T. Kuballa, M. Ohmenhaeuser, S. Rohn, P. Weller, Differentiation of the botanical origin of honeys by fast, non-targeted 1H-NMR profiling and chemometric tools as alternative authenticity screening tool. *Proceedings of the XIII International Conference on the Applications of Magnetic Resonance in Food Science*, Karlsruhe, Germany, 7–10 (2016).