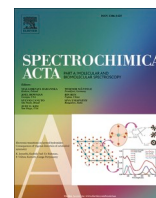




Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Comparing post-acquisition methods for Raman-based mineral identification of particles in yellow earth pigments

Marcus A. Johns^a, Frederik Handberg^b, Hendrik Groß^c, Teodora Boyadzhieva^d,
Ayoub Laghrissi^a, Thea Zeeck^c, Serkan Ayvaz^b, Casper Kunstmann-Olsen^a,
Horst-Günter Rubahn^a, Lorenz Kienle^c, Jacek Fiutowski^{a,*}

^a Mads Clausen Institute, NanoSYD, University of Southern Denmark, Alsion 2, Sønderborg 6400, Denmark

^b Centre for Industrial Software, University of Southern Denmark, Alsion 2, Sønderborg 6400, Denmark

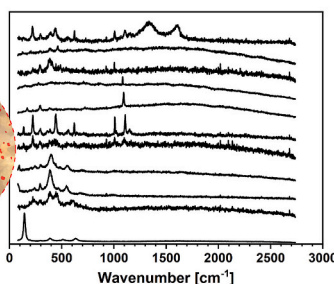
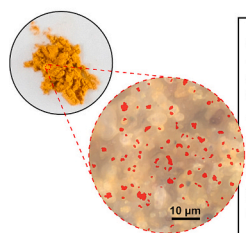
^c Kiel University, Department for Materials Science, Kaiserstr. 2, 24145 Kiel, Germany

^d Conservation Centre Vejle, Maribovej 10, Vejle 7100, Denmark

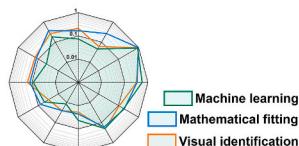
GRAPHICAL ABSTRACT

- Particle-resolved Raman enables robust, non-destructive pigment differentiation.
- Combining Raman with statistical analysis reveals key mineral markers of provenance.
- Visual identification is flexible but slow and less reliable for subtle differences.
- Machine learning and mathematical fitting are consistent but need reference spectra.
- Semi-automated methods distinguish all five yellow earth pigments considered.

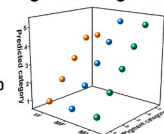
Particle mineral spectra



Determined mineral content



Pigment assignment



ARTICLE INFO

Keywords:

Earth pigments
Confocal Raman spectroscopy
Chemometrics
Machine learning

ABSTRACT

Natural yellow earth pigments are heterogeneous mineral assemblages dominated by iron (hydr)oxides and accompanied by variable accessory phases such as quartz, calcite, gypsum, and aluminosilicates. Although these accessory minerals may encode geological origin, their reliable identification from bulk Raman spectra is complicated by multiphase band overlap, fluorescence background, and particle-level heterogeneity. As a result, mineral discrimination from Raman data alone remains analytically demanding. Here, we examine whether

* Corresponding author.

E-mail address: fiutowski@mci.sdu.dk (J. Fiutowski).

<https://doi.org/10.1016/j.saa.2026.128238>

Received 13 April 2026; Received in revised form 15 May 2026; Accepted 9 June 2026

Available online 10 June 2026

1386-1425/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

systematic post-acquisition processing can enable robust mineral-based differentiation of closely related pigments using Raman spectroscopy alone. Five yellow earth pigments (three yellow ochres and two raw siennas) were analyzed using particle-resolved confocal Raman spectroscopy, generating statistically representative datasets (>200 spectra per sample). Mineral identification was performed using three strategies, visual interpretation, automated three-component spectral library fitting (3C model), and supervised machine learning classification (XGBoost), and assessed using a binary presence/absence scheme to support comparison across heterogeneous particle populations. Both the 3C and XGBoost approaches enabled statistical separation of all five pigments, whereas K-means cluster analysis of the visual interpretation did not distinguish between the two raw siennas. Together, these results support a reproducible Raman-only workflow for preliminary mineral assignment and pigment discrimination in provenance-oriented cultural heritage studies.

1. Introduction

Yellow earth pigments have been used for millennia in artistic, decorative, and utilitarian contexts, prized for their warm hues, natural abundance, and relative stability [1,2]. These pigments have historically been produced from earth deposits rich in iron (III) hydroxides, namely goethite ($\alpha\text{-FeO}(\text{OH})$), that also contained iron (III) oxides, such as haematite (rhombohedral Fe_2O_3), as well as other iron (III) complexes, such as jarosite ($\text{KFe}_3(\text{SO}_4)_2(\text{OH})_6$). Additional accessory minerals, such as aluminosilicates, quartz (SiO_2), calcite (CaCO_3), gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) and anatase (tetragonal TiO_2) may also be present, which are not only intrinsically linked to the earth pigment colour but are also dependent on the geological origin [3,4].

While iron (hydr)oxides largely determine the visual appearance of yellow earth pigments, it is often the accessory mineral phases and subtle chemical substitutions within dominant phases that encode deposit-specific formation conditions. Identifying the origin of the pigment is of particular interest to the field of heritage science and conservation because it enables researchers to reconstruct population mobility. It allows them to determine whether materials were taken from the local area, or whether there were trade/migration routes from places further afield at the time of the object under analysis [5,6].

Typically, multiple analytical techniques are used to fully characterize the pigment in question [7,8]. These often include colorimetry, scanning electron microscopy (SEM), X-ray diffraction (XRD), X-ray fluorescence spectroscopy (XRF), energy-dispersive X-ray analysis (EDX/EDS), laser-induced breakdown spectroscopy (LIBS), and vibrational spectroscopies. Whilst manual analysis of the resulting data sets is feasible [5,9,10], chemometrics, the application of multivariate statistical analysis to extract information from complex data sets, is preferred as it enables semi-automation of the process [11]. At the simplest level, this involves manual comparison of multivariate analytics, such as principal component analysis (PCA), to establish differences between samples [12,13]. However, the fusion of features from both elemental and molecular data sets prior to analysis typically improves identification. For example, both Ramos et al. and Donais et al. demonstrated that combining XRF (elemental analysis) and Raman (molecular analysis) spectroscopies enabled differentiation of earth pigments, which was not possible based on spectral data from a single technique [14,15]. Similarly, the fusion of laser-induced breakdown spectroscopy (elemental analysis) with Raman spectroscopy has been successfully applied to classify minerals on both Earth and Mars [16,17]. However, implementing multi-technique analysis requires access to suitable instrumentation and a guarantee of spatial correlation between techniques. The latter can be particularly challenging for high-resolution analysis or when specific sample preparation is required, such as applying a conductive coating before SEM-EDS. Therefore, the development of post-acquisition analytical processing that enables distinction among similar pigments using a single technique would be particularly beneficial for preliminary studies or when access to multiple complementary instruments is limited.

Vibrational spectroscopies, including Fourier transform infrared spectroscopy (FTIR) and Raman spectroscopy, are favored within the community because they offer high molecular specificity whilst being

non-destructive and non-invasive, enabling analysis both in the laboratory and on-site with minimal sample preparation. Whilst background fluorescence can be problematic, Raman spectroscopy typically offers better spatial resolution than FTIR. As such, multiple Raman pigment and mineral databases are available in the literature, alongside numerous examples of the application of PCA to Raman data for cultural heritage analysis [18–26].

Despite the availability of extensive Raman reference libraries, consistent identification of mineral phases in heterogeneous, multiphase earth pigments remains analytically challenging when relying on a single technique. As vibrational spectroscopies provide molecular-specific information, changes in the mineral content of earth pigments can be detected from their spectra [4]. This is most readily observed by comparing spectra reported in databases. For example, for yellow ochre, Bell et al. reported bands at ca. 416 and 1008 cm^{-1} [18], which could be attributed to gypsum [25], whilst Marucci et al. reported bands at ca. 617, 1005 and 1134 cm^{-1} , also potentially gypsum, and one at ca. 1084 cm^{-1} [26], which could be attributed to calcite [27]. Similarly, the bands at ca. 1012 and 1109 cm^{-1} for raw sienna reported by Bikiaris et al. could be assigned to jarosite [19,25]. Raman bands reported in the literature for yellow ochre and raw sienna pigments are available in the Supplementary Material (Tables S1, S2).

Typically, multivariate analysis is applied to distinguish bulk spectra from one another across their entire spectral range rather than to identify the content of specific minerals within each spectrum. This is because the relative band intensities cannot be used to quantify the amount of a given mineral, as spectral intensities are not directly comparable between measurements. However, the contribution of minor mineral components may be completely overlooked due to their negligible spectral intensity in the bulk material, making it more challenging to distinguish earth pigments. The approach also assumes that the spectra acquired are representative of a homogeneous sample, which may not be the case for these multi-component pigments, especially at high spatial resolution.

In heterogeneous materials such as natural earth pigments, discrimination depends less on subtle intensity differences in averaged spectra and more on the consistent identification of mineral fingerprints across individual particles. Accordingly, comparing mineral occurrence statistically across spectra is essential for distinguishing closely related pigments. Therefore, the ability to compare statistical differences in the mineral content of individual pigments is required to overcome these challenges. For this approach, the fingerprint spectrum for each mineral needs to be consistently identified under challenging conditions, e.g., band overlap from different minerals and fluorescence background. Various methods have been applied to address this, including manual visual identification, multi-component mathematical fitting, such as the TrueMatch software developed by WITec, and machine learning techniques [28–30]. Each of these approaches embodies different assumptions regarding how spectra should be interpreted and may therefore differ in robustness, reproducibility, and susceptibility to bias.

In this study, we address whether post-acquisition analytical processing of Raman spectra alone can reliably distinguish between closely related earth pigments and whether the choice of analytical approach affects effectiveness. We consider the identification of five yellow earth

pigments (three yellow ochre and two raw sienna) based on their relative mineral contents via high-resolution, particle-specific Raman spectroscopy. Rather than relying on intensity-based quantification—an approach that assumes comparability of signal strength across heterogeneous particles—we recast mineral analysis as an occurrence problem: for each spectrum, we record whether a detectable fingerprint of a given mineral is present, independent of signal strength.

Three alternative methods are applied for post-acquisition analysis of mineral spectra: manual visual assessment, mathematical modelling using commercially available software (3G model), and assisted machine learning (XGBoost). Furthermore, we address the challenge of quantitative mineral analysis by only considering whether a given spectrum contains a detectable signal for a given mineral and then normalizing the total number of counts per mineral in a sample to the counts of the mineral that was most common to that sample. This normalized occurrence-based framework yields comparable mineral profiles that are robust to differences in sampling density and spectra classified as noise. We demonstrate that this novel approach, combined with PCA, enables statistically reproducible discrimination among the five pigments using both mathematical fitting and assisted machine learning for mineral identification.

2. Materials & methods

2.1. Materials

Two yellow ochre samples from Andalusia, Spain (YOA, product no. 11272) and Burgundy, France (YOB, product no. 11573), as well as one raw sienna sample from Italy (RSI, product no. 40400), were obtained from Kremer Pigmente GmbH & Co. (Aichstetten, Germany). One yellow ochre sample from Gargas, France (YOG, Ocre jaune foncé JFLES) was acquired from Société des Ogres de France (Vaucluse, France) and one raw sienna sample of unknown provenance (RSD) from J.C. Juul Eff. (Aarhus, Denmark). Information on the mineral content of each pigment was not provided by the manufacturers, although Société des Ogres de France stated that YOG consisted of kaolinite and goethite.

2.2. Digital microscopy & laser-induced breakdown spectroscopy (LIBS)

Digital microscopy and LIBS were performed on a Keyence (Frankfurt, Germany) VHX-X1, equipped with an EA-300 material analysis unit (laser: 355 nm Nd:YAG, spot size: 10 μm), under ambient conditions. After imaging, samples were compacted prior to elemental analysis over a $750 \times 750 \mu\text{m}^2$ grid ($N = 3$, $n = 25$).

2.3. Scanning electron microscopy (SEM) & energy dispersive X-ray spectroscopy (EDS)

Pigment samples were coated with 5 nm Au prior to SEM-EDS characterization to reduce charge build-up. SEM was performed on a Hitachi (Marunouchi, Japan) S-4800 high-resolution scanning electron microscope (secondary emission, accelerating voltage: 15 kV) equipped with a Bruker (Billerica, MA) XFlash 6 EDS system. Elemental analysis of each sample was performed at a magnification of $5000\times$ at five separate points ($N = 1$, $n = 5$), whilst elemental mapping was performed at a magnification of $1000\times$.

2.4. Transmission electron microscopy (TEM) & EDS

TEM was performed on a probe-corrected JEOL (Akishima, Japan) JEM-ARM200F NEOARM scanning transmission electron microscope, operated at 200 kV (cold-FEG) and equipped with an EDS system with a dual silicon drift detector. Simulated electron diffraction patterns were created using the software jems v. 4.13531u2024b31 by P. Stadelmann.

2.5. Raman spectroscopy

Raman spectroscopy was performed on a WITec (Ulm, Germany) Alpha 300 confocal microscope (laser excitation wavelength: 633 nm; laser power: 1 mW; objective: $100\times$; spot size: ca. 1 μm ; integration time: 10 s; accumulations per spectrum: 10; wavenumber range: $80\text{--}2735 \text{ cm}^{-1}$). To automatically generate sufficient spectra for analysis, the WITec ParticleScout software was used. Briefly, threshold analysis was utilized to generate points for Raman analysis from a single image (watershed analysis, bright spots, background threshold: 0, detection threshold: 0). The total number of spectra per image was limited to between 200 and 250 by adjusting the minimum area (0–0.2) and circularity (0–0.5) values during threshold analysis. This was sufficient to be at least 95% confident that the observed mineral fraction was representative of the area analyzed for each pigment. A spectral z-scan prior to the acquisition of each spectrum was utilized to optimize focus for spectral acquisition (integration time: 5 s, step size multiplier: 5). Complete analysis was performed on three sample images per pigment ($N = 3$, $n \geq 200$) (Supplementary Material, Table S3). Representative z-stack images of the pigments used for threshold analysis and an example of the areas identified within an image via threshold analysis are available in the Supplementary Material (Fig. S1, S2).

2.6. Mineral identification from Raman spectra

Minerals were manually identified by visually comparing obtained spectra to those reported in the literature [18–22,25–27]. Manual deconvolution of spectra that matched those in the literature was performed using Fityk, an open-source curve-fitting program [31], to generate ideal spectra with no background signal (Fig. 1c). The band center wavenumbers were then compared to those in the literature to ensure that the generated ideal spectra were representative of the minerals (Supplementary Material, Tables S4–S13). Slight shifts in the deconvoluted band wavenumbers and the values reported in the literature may be due to the deconvolution procedure as the fitted bands do not necessarily align with the raw data maxima.

As Raman signal intensities cannot be used to quantitatively assess mineral content, a simple binary presence-absence approach was applied, in which the detection of a characteristic mineral Raman fingerprint in the spectrum was counted once, irrespective of signal strength. For each image, mineral detections were scaled relative to the most frequently detected mineral, allowing samples with different numbers of measured points or spectra with no discernible signal to be compared. Spectra without identifiable mineral signals were classified as ‘noise’ and excluded from normalization.

2.7. Visual assessment of Raman spectra

Minerals were manually identified by comparing the sample spectra to those of the mineral spectra and ensuring that a) wavenumbers matched and b) band height ratios were visually similar. Due to the targeting of individual particles at high magnification, the fingerprints for individual minerals were readily identifiable within the Raman spectra (Supplementary Material, Fig. S3–S12), unlike conventional bulk analysis where careful deconvolution is required. The authors recognize that identification should typically only be considered positive if multiple bands are identified. However, in this instance, for minerals with spectra consisting of a dominant band and additional, less prominent bands (quartz, gypsum, calcium carbonate, ankerite, anatase), the identification of the single dominant band was considered sufficient for a positive count.

2.8. Automated three-component spectral library fitting of Raman spectra (3C model)

Spectral analysis was automated using the TrueMatch software

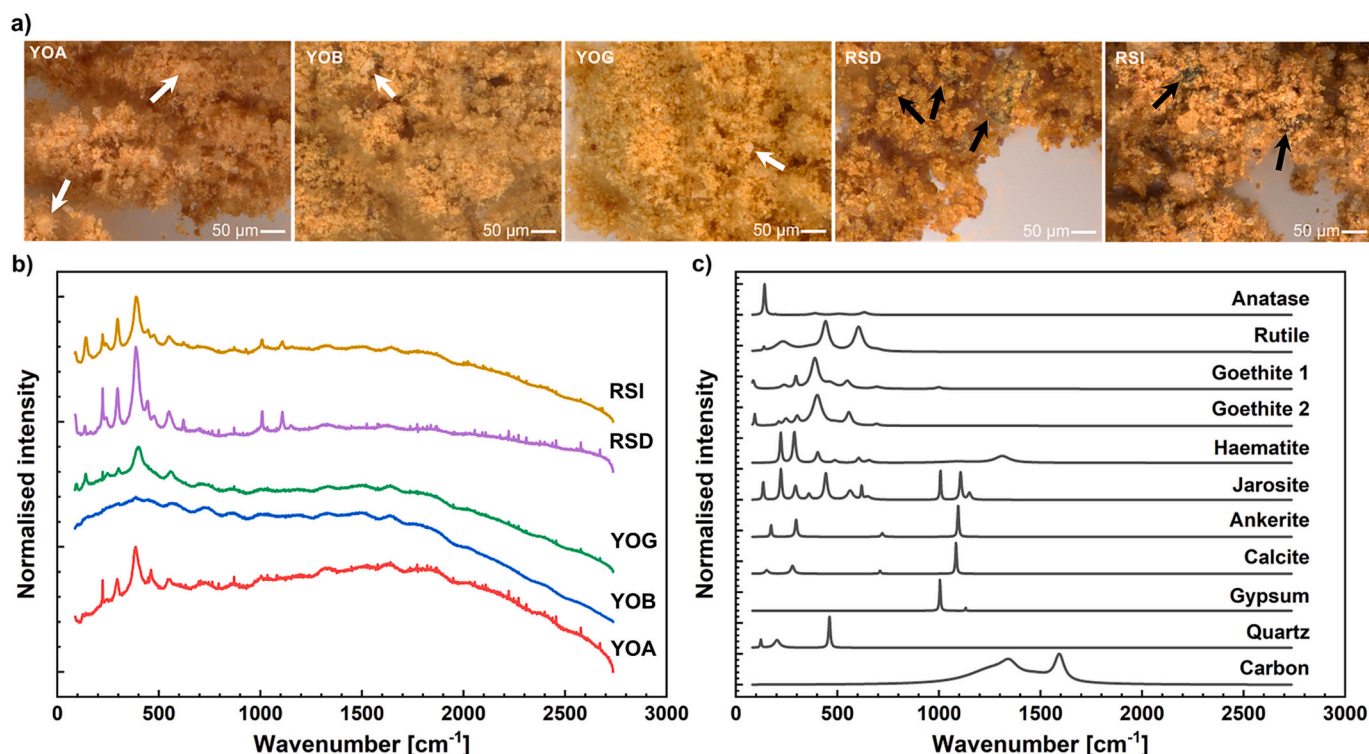


Fig. 1. a) Digital microscope images of pigments (white arrows: larger, colorless particles, likely quartz; black arrows: larger, dark particles, likely iron-based); b) Offset normalized average spectra for pigments; c) Fitted Raman spectra of minerals identified in yellow earth pigments. Examples of raw spectra for each mineral identified in each pigment provided in Supplementary Material (Fig. S5-S12) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(WITec, Ulm, Germany). A three-component search was utilized with the database consisting of the ideal spectra of the previously identified minerals. The result with the highest hit quality index (HQI), which was determined for multiple models via correlation coefficient analysis, was automatically considered optimal, and, based on initial observations, spectra with an HQI less than 25% were treated as noise. The software always assigned three components to each spectrum, resulting in negative component fractions for spectra containing fewer than three components. In these instances, the minerals identified with negative fractions were ignored.

2.9. Supervised machine learning analysis of Raman spectra (XGBoost)

Supervised machine learning analysis of the spectra was performed using an XGBoost model. In brief, the model consisted of three steps:

Data generation: 500 augmented samples were generated for each non-synthetic mineral reference spectrum. Augmentation included intensity scaling (0.8–1.2×), noise injection (Gaussian, Poisson, and salt-and-pepper), peak shifting (± 2.0 cm⁻¹), background modelling using polynomial and sinusoidal functions, modulation of peak width and height, and synthetic signal composition to simulate spectra from mixed mineral samples. Data sets for each mineral contained equal numbers of signal/no-signal samples with ground truth labels manually assigned.

Data preprocessing: All spectra underwent a five-stage preprocessing pipeline (wavenumber filtering, Savitzky-Golay smoothing, asymmetric least squares baseline correction, L2 normalization, spectral fingerprint extraction utilizing dynamic window sizes).

Mineral classification: Independent binary (signal/no-signal) classifiers were developed for each mineral. Hyperparameter tuning was performed using Bayesian optimization. Each dataset was split into training and testing subsets with an 80:20 ratio with stratified sampling. Probability scores (0.0–1.0) for each mineral spectrum were generated, which were converted into the binary classifiers using a threshold of 0.5.

The trained model was then applied to the sample spectra with the presence of each mineral assessed independently from the others.

Complete details, including a schematic of the workflow and mineral confusion matrices, are available in the Supplementary Material (Fig. S13, S14).

2.10. Statistical analysis

Statistical analysis was performed for each post-acquisition analytical method using IBM SPSS Statistics software. A one-way analysis of variance (ANOVA) was used to determine statistical differences among the means of three samples, each containing more than 200 spectra ($N = 3$, $n \geq 200$) per pigment, assuming equal variances. Differences were considered significant with a confidence level of 0.95 (95% confidence). Bonferroni post hoc comparisons were used to group pigments based on statistically insignificant differences for each mineral component.

2.11. K-means cluster analysis

K-means cluster analysis was performed for each post-acquisition analytical method using IBM SPSS Statistics software (Armonk, NY). Briefly, the mineral content and noise score factors for each sample were standardized to the mean value for the factor (z-scores). Principal component analysis (PCA) was applied to the z-scores (12 factors) and principal components (PCs) with Eigenvalues greater than one. K-means cluster analysis was then performed on these PCs. Silhouette analysis was performed to optimize the number of clusters for each analytical method based on maximizing the average silhouette score whilst ensuring that no clusters consisting of a single value were formed.

3. Results & discussion

3.1. Initial mineral analysis

Samples of the same group (yellow ochre or raw sienna) were comparable in appearance (Fig. 1a). The ochres consisted of a homogeneous structure with similar grain sizes. All samples included colorless, transparent particles of slightly larger grain sizes, most likely quartz, which were most apparent in YOA (Fig. 1a, white arrows). In comparison, the raw sienna samples were significantly rougher and inhomogeneous, and included dark-colored grains, indicating an iron-rich component such as haematite, jarosite or goethite (Fig. 1a, black arrows).

Eleven unique minerals were identified in the Raman spectra from the five pigments (Fig. 1b,c), which were classified into four groups based on common metallic elements (Table 1). Elemental analysis via LIBS and SEM-EDS confirmed the presence of the elements identified in these compounds (Supplementary Material, Table S14, S15). Whilst the absolute elemental content of the samples as determined by each technique varied, the common trends for observed for the primary components: Si and Al content was higher in the yellow ochres, whilst Fe content was higher for the raw siennas. Variances may have been due to differences in magnification used by the two techniques and the heterogeneity of the samples, e.g., particle size and distribution specific to each mineral, alongside the use of different subsamples for each technique. Ternary plots of the elemental fractions provide further information on the likely minerals present within each sample (Fig. 2). Aluminosilicates dominated the YOB and YOG samples, which may explain the relatively poor Raman signal for these samples as aluminosilicates were not detected under the Raman conditions used (Fig. 1b). The distributions indicated a high albite/orthoclase/quartz content in YOB with a minor kaolinite/anorthite component and very low fraction of Fe-containing minerals, whilst YOG had a dominant kaolinite/quartz fraction. YOA appeared to be predominantly quartz with a minor kaolinite fraction, the broader distribution potentially indicative of a wider range of minerals present. In contrast, the raw siennas were dominated by Fe-containing fractions, with RSI potentially having a greater quartz/orthoclase content than RSD, resulting in a lower overall Raman signal intensity (Fig. 1b). The normalized quartz count for each pigment, as determined by Raman spectroscopy and averaged over the three post-acquisition methods, strongly correlated with the corresponding Si:Fe content, as determined by LIBS elemental analysis (Supplementary Material, Fig. S20). This correlation indicated that the mineral counts determined using the Raman analytical methodology were representative of the true mineral content for the detectable minerals. Bulk ATR-FTIR spectra were also obtained for each pigment, which indicated the presence of the minerals identified via Raman spectroscopy as well as hydrated aluminosilicates, which were particularly prominent in YOB and YOG, in agreement with the SEM-EDS analysis (Supplementary Material, Fig. S21).

Table 1

Unique minerals identified in the 633 nm Raman spectra of the pigment samples grouped by common metallic element.

Titanium	Iron	Calcium	Other
Anatase (TiO ₂)	Goethite 1 (α -FeO(OH))	Ankerite	Quartz (SiO ₂)
Rutile (TiO ₂)	Goethite 2 (α -(Fe,Al)O(OH))	Calcite (CaCO ₃)	Carbon (C)
	Haematite (Fe ₂ O ₃)	Gypsum (CaSO ₄ ·2H ₂ O)	
	Jarosite (KFe ₃ (SO ₄) ₂ (OH) ₆)		
	Ankerite (Ca(Fe,Mg,Mn)(CO ₃) ₂)		

3.2. Specific analytical features

Four features associated with the analysis should be noted: Firstly, two distinct Raman spectra that shared features associated with reported spectra for goethite were consistently identified (Supplementary Material, Table S6). Neither spectrum matched the spectra for common FeO(OH) polymorphs, such as akageneite (β -FeO(OH)), feroxyhyte (δ -FeO(OH)), and lepidocrocite (γ -FeO(OH)) [32], nor to an alternative mineral in the RRUFF database [25]. However, Wu et al. reported a shift in the primary band associated with goethite from 387 cm⁻¹ to 398 cm⁻¹ upon incorporation of Al, i.e., from α -FeO(OH) to α -(Fe, Al)O(OH) [33]. Higher counts of the goethite 2 spectrum (primary band: 400 cm⁻¹) were identified in YOB and YOG (Fig. 4, Supplementary Material, Fig. S5), which also exhibited higher Al contents. STEM-EDS analysis of YOG confirmed an Al content of 7 at.% in particles that exhibited the needle-like crystal structure of goethite (Fig. 3a), indicating a chemical formula of α -(Fe_{0.79}Al_{0.21})O(OH). In a further analysis by high-resolution TEM combined with electron diffraction (ED), the experimental data was compared to a simulated pattern of highly Al-substituted α -(Fe_{0.83}Al_{0.17})O(OH) [34] in zone axis [312] (Fig. 3b), exhibiting a good correlation.

Secondly, the extracted spectrum of gypsum consisted of only two bands at ca. 1004 and 1131 cm⁻¹. However, additional bands should be readily identified at ca. 414, 493 and 619 cm⁻¹ (Supplementary Material, Table S11). A spectrum consisting solely of gypsum was not observed during the analysis of the pigments. As a result, the three additional bands were indistinguishable from those in the other minerals, negatively affecting the extracted spectrum. Thirdly, it must be considered that the likelihood of mineral detection will be dependent on the effectiveness of the mineral to scatter the Raman signal. For example, TiO₂ phases are better Raman scatterers than iron oxides, increasing their likelihood of being detected relative to their actual content, whilst aluminosilicates are evidently poorer. Finally, manganese, typically present in raw sienna pigments as manganese oxide, was only detected in trace amounts (0.4 at.%) by SEM-EDS in the RSI sample (Supplementary Material, Table S15).

3.3. Statistical analysis of the three post-acquisition approaches

Significant differences between the five pigments were observed in all three post-acquisition analyses of the Raman data (Fig. 4). As expected, goethite dominated the spectra, with type 1 being the most common mineral component for the YOA, YOB, RSD and RSI pigments and the aluminium-substituted type 2 being the most common for the YOG pigment (Fig. 4a). Type 2 was also prominent in the YOB sample in all three analyses, which enabled distinction between the three yellow ochre pigments (Fig. 4b). The yellow ochre and raw sienna pigments were readily distinguished from one another by their jarosite content, with significantly higher counts present in the raw sienna pigments in agreement with the SEM-EDS elemental distributions (Fig. 2) and increase in increase in FTIR transmission intensity around 1090 cm⁻¹ (Supplementary Material, Fig. S21).

Significant differences were also observed in the accessory minerals, although this was dependent on the type of analysis used. Regarding visual identification, the yellow ochre samples differed statistically in the content of two minerals, anatase and quartz (Fig. 4bi). Anatase was most common in YOB, whilst quartz was most prominent in YOA. This was supported by the elemental analysis – LIBS detected traces of Ti in YOB, and the Si content was highest in YOA. However, these statistical differences were not observed to the same extent if analyzed via the 3C model or XGBoost. The YOA pigment statistically had a higher quartz content and lower anatase content compared to YOB and YOG, but no statistical difference was observed between the YOB and YOG samples (Fig. 4bii,iii). One potential explanation is that unconscious bias may have influenced the visual identification of minerals. The user may have been more likely to positively identify a mineral in borderline cases

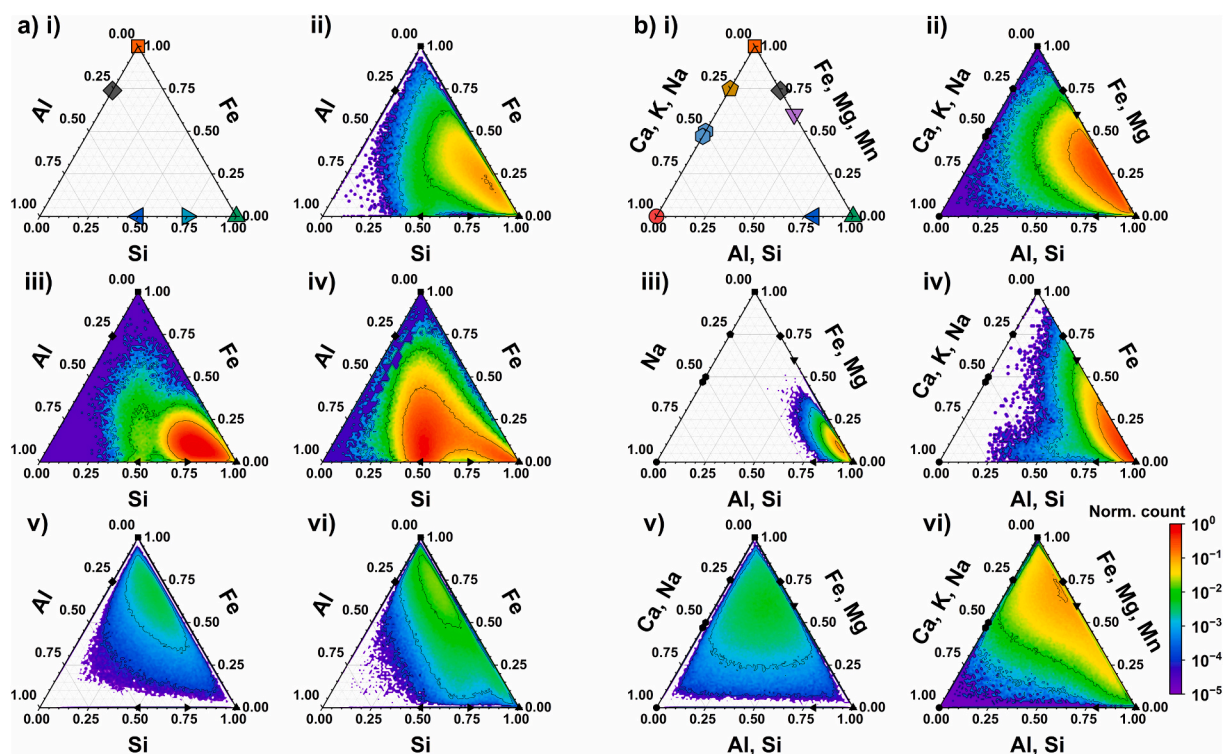


Fig. 2. a) Ternary plots of Al, Fe and Si fractions (at.%) for i) ideal minerals (orange square: Fe_1 , goethite, haematite, ankerite, magnetite, jarosite; black diamond: $\text{Al}_{0.26}\text{Fe}_{0.74}$, Al-substituted goethite and haematite; green triangle: Si_1 , quartz, serpentines; cyan right triangle: $\text{Al}_{0.25}\text{Si}_{0.75}$, albite, orthoclase; blue left triangle: $\text{Al}_{0.5}\text{Si}_{0.5}$, anorthite, kaolinite); ii-vi) maps of normalized SEM-EDS micrograph pixel count of fractions (at.%) for ii) YOA, iii) YOB, iv) YOG, v) RSD, and vi) RSI.; b) Ternary plots of (Al, Si), (Ca, K, Na) and (Fe, Mg, Mn) fractions (at.%) for i) ideal minerals (orange square: $(\text{Fe,Mg})_1$, goethite, haematite, glushinskite, magnetite; black diamond: $(\text{Al,Si})_{0.26}(\text{Fe,Mg})_{0.74}$, Al-substituted goethite and haematite; purple inverted triangle: $(\text{Al,Si})_{0.4}(\text{Fe,Mg})_{0.6}$, serpentines; green triangle: $(\text{Al,Si})_1$, quartz, kaolinite; blue left triangle: $(\text{Al,Si})_{0.8}(\text{Ca,K,Na})_{0.2}$, albite, anorthite, orthoclase; red circle: $(\text{Ca,K,Na})_1$, calcite, gypsum; light blue hexagon: $(\text{Ca,K,Na})_{0.5}(\text{Fe,Mg})_{0.5}$, ankerite; yellow pentagon: $(\text{Ca,K,Na})_{0.25}(\text{Fe,Mg})_{0.75}$, jarosite); ii-vi) maps of normalized SEM-EDS micrograph pixel count of fractions (at.%) for ii) YOA, iii) YOB, iv) YOG, v) RSD, and vi) RSI. Black shapes correspond to position of ideal minerals; fractions rounded to 2 d.p. ($N = 1$, $n = 905,216$). SEM micrographs and elemental maps available in Supplementary Material (Fig. S15-S19) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where the signal may have been noise if they believed that the mineral was common to that pigment and vice versa. This would not have been possible in the 3C model and XGBoost analyses.

Visual identification allowed statistical separation of the raw sienna pigments based on differences in their anatase content only, as was also the case for the 3C model and XGBoost analyses, despite Ti only being detected in trace amounts for both pigments via SEM-EDS (Supplementary Material, Table S15). While anatase emerged as a common discriminator across all the methods considered, the 3C model revealed further significant differences in haematite and ankerite content, whereas the XGBoost model identified gypsum and carbon as additional distinguishing components (Fig. 4b). Possible reasons for why the statistical analyses were not consistent between analytical methods are discussed below. It was also noted that spectra were more likely to be classified as noise in the YOB pigment than in the other four pigments across all three methods due to the higher aluminosilicate content. It is apparent from the average spectra (Fig. 1b) that whilst the backgrounds for YOA, YOG, RSD and RSI are slightly different from one another, signals from the mineral components present are visually evident, which was not the case for YOB.

3.4. Comparison of the post-acquisition approaches

Variations in the average relative component contents were observed dependent on the analytical method (Fig. 5a; Supplementary Material, Fig. S22). The greatest variation from the average value of the three methods for each component was observed for two components (haematite and quartz) for visual identification; four for the 3C model

(rutile, jarosite, gypsum and noise); and six for the XGBoost model (anatase, goethite types 1 and 2, ankerite, calcite and carbon). These were generally within an acceptable range, defined as plus/minus one count per hundred counts of the dominant mineral component (Fig. 5b-d). For both visual identification and XGBoost, there were instances in which these boundaries were exceeded, involving a single mineral and one or two pigments (Fig. 5; Supplementary Material, Fig. S23, S24). Examples included underestimation of jarosite in YOG for visual identification and XGBoost overestimation of goethite type 2 for the raw sienna samples. As the same minerals fell within the acceptable variation range for the other pigments, it was considered that the apparent error was either due to user misidentification or to spectral processing prior to analysis, rather than an inherent error in the approaches themselves. However, in the case of the 3C model, rutile was overestimated in every pigment compared to the other two approaches (Fig. 5; Supplementary Material, Fig. S25). Given that the rutile bands are relatively broad and overlap with the bands of the dominant minerals (Fig. 1c), we hypothesize that its inclusion improved the overall correlation coefficient to the spectra rather than due to the bands being identifiable. This was not an issue for the visual identification and XGBoost approaches, as pigment assignment was primarily based on band identification relative to the background. One further concern with the 3C model was that the software limitation of three components per spectrum would potentially result in the underestimation of minerals. However, this was not observed despite a maximum of six mineral components being identified in a single spectrum for both the visual and XGBoost analyses. This was likely due to the low fraction (2.6% for visual identification, 2.9% for XGBoost) of mineral-containing spectra that

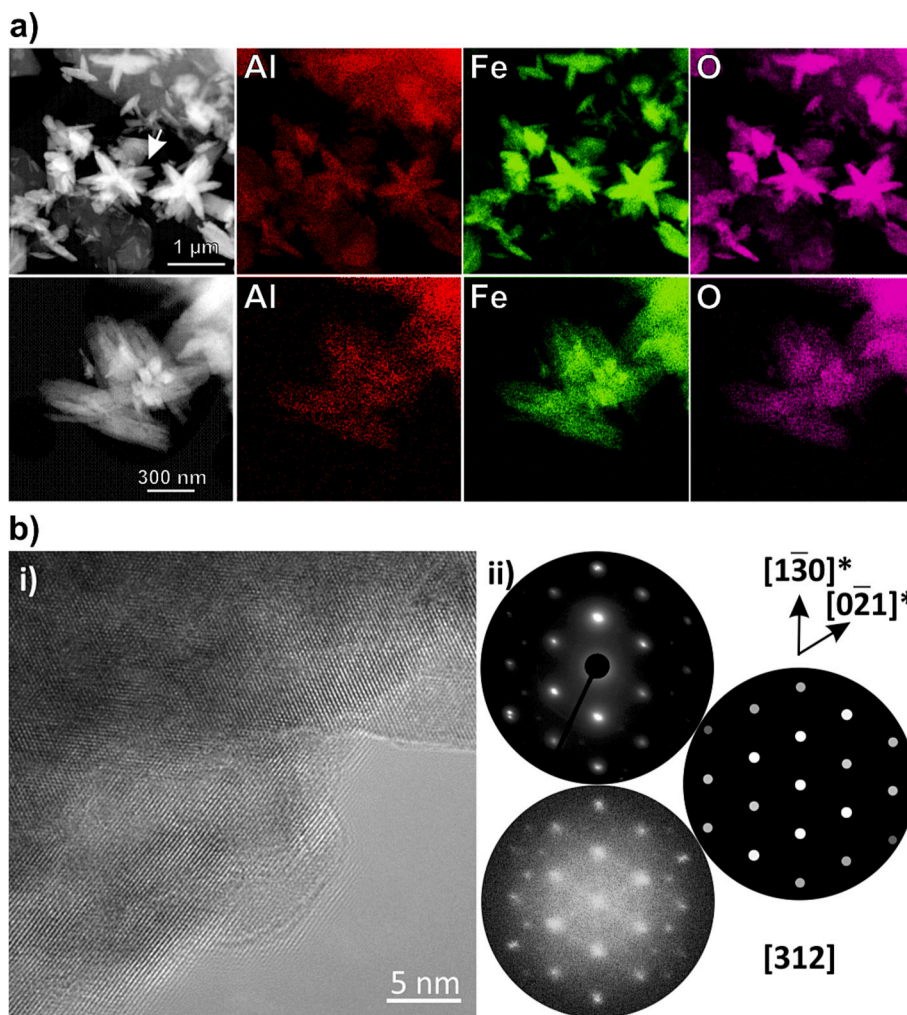


Fig. 3. a) STEM micrographs and EDS elemental maps of particles identified as Al-containing goethite from YOG (white arrow: particle investigated by HRTEM); b) i) HRTEM micrograph of the edge of a particle identified as Al-containing goethite from YOG; ii) Corresponding electron diffraction pattern (top left), FFT (bottom left) and simulated diffraction pattern in zone axis $[312]$ (right). The simulation is based on $\alpha\text{-(Fe}_{0.83}\text{Al}_{0.17}\text{)O(OH)}$ (5.7 at.% Al substitution) [34].

contained more than three components (Supplementary Material, Tables S16, S17). It is also worth noting that both the 3C model and XGBoost approaches had thresholds for positive identification, which were arbitrarily set by the authors ($\geq 25\%$ HQI for 3C and ≥ 0.5 confidence level per component for XGBoost). Fine-tuning of these values may further improve these models, but this was not considered in this analysis.

The differences in mineral component fractions between each analytical method are likely to be due to the different approaches taken for mineral identification, which also impact on their relative advantages and disadvantages (Table 2). At one ‘extreme’, the 3C model aims to minimize the correlation coefficient between the model and the data. It accurately deconvolutes overlapping bands from one another but has a higher risk of overfitting to noise, as observed in the case of the over assignment of rutile. At the other ‘extreme’, XGBoost is trained to distinguish the Raman bands for specified components from the background noise. This approach is closer to manual analysis, but the training model must consider all spectral scenarios that may be present in a sample. For example, if the bands of two components overlap, one component may not be identified unless that scenario has been included in the training data. Although, it is noted that alternative machine learning approaches may mitigate this [35]. Both models also require a database of representative component spectra prior to analysis, which is not the case for visual identification by the user so long as they have access to suitable literature resources. Visual identification primarily

focuses on band combinations, as with the XGBoost model, but, dependent on the user's expertise, can also include identifying changes in band shape or the presence of band shoulders to identify multiple components, which is closer to the 3C model. However, visual analysis is a lengthy process, is potentially susceptible to unconscious bias, and may find low-signal component identification more challenging than other approaches. The authors note that visual identification resulted in the fewest spectra being classified as lacking identifiable mineral components compared to the other methods (Supplementary Material, Fig. S26). However, no significant differences between the fraction of spectra assigned as noise for each method were observed for the pigment samples analyzed, suggesting that this did not have a critical impact on the analysis. XGBoost was better at mineral assignment for the challenging sample YOB, albeit potentially at the expense of a higher component misidentification rate (Supplementary Material, Fig. S22b, Table S16). One possible explanation is that the background-subtraction method used introduces uncertainty into the data, such as high background noise levels or the presence of unidentifiable or false ‘bands’ (Supplementary Material, Fig. S27), which negatively impacts both the 3C and XGBoost approaches.

3.5. Suitability of the approaches for pigment identification

Principal component analysis of the pigments identified three PCs for the visual identification approach and four PCs for both the 3C and

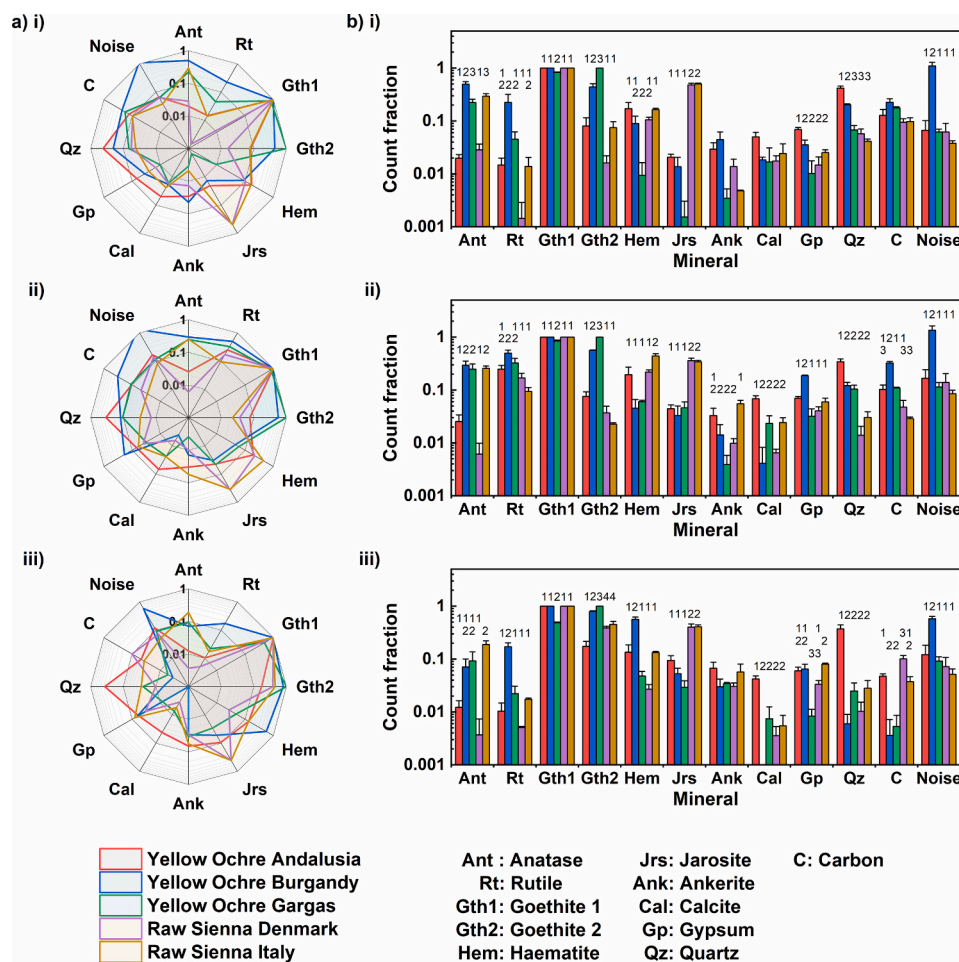


Fig. 4. a) Radar plots of the relative component content for each pigment; b) Statistical comparison of relative counts of each component between each pigment. i) Visual identification; ii) 3C model; iii) XGBoost model. Numbers indicate pigment groups that are statistically different from one another for a given component. $N = 3$, $n \geq 200$, error bars: S.E, $p < 0.05$.

XGBoost approaches (Supplementary Material, Fig. S28-S31). PCs 1–3 separated the three yellow ochre samples from one another and the raw sienna samples, whilst PC 4 enabled separation of the two raw sienna samples (Fig. 6; Supplementary Material, Table S18-S20). Analysis of the PC factor scores revealed common mineral component features to each PC across all three post-acquisition analyses (Supplementary Material, Fig. S29-S31, Table S21). In brief, PC1 separated YOB based on the higher noise and apparent rutile fractions, which agreed with Ti only being detected in the YOB sample via LIBS analysis (Supplementary Material, Table S14); PC2 separated YOG based on the goethite type 1:2 (α -FeO(OH): α -(Fe,Al)O(OH)) ratio; PC3 separated YOA based on the higher calcite and quartz contents, in agreement with elemental analysis (Supplementary Material, Tables S14, S15); and PC4 separated RSI from RSD based on the higher anatase content. K-means cluster and silhouette analysis confirmed that all five pigments could be identified separately from one another based on the 3C and XGBoost approaches (Fig. 6b,c; Supplementary Material, Tables S22-S25). The XGBoost approach resulted in a slight improvement in the separation of the pigments over the 3C model, indicated by a higher average silhouette score and smaller standard deviation (Supplementary Material, Table S25). However, visual identification, which had one less PC, was unable to separate the two raw sienna samples from one another, which were also confounded with one of the YOB samples (Fig. 6a). Thus, greater consistency in mineral assignment across samples by the semi-automated techniques may be critical for ensuring accurate multi-mineral pigment identification.

4. Conclusions

Accessory mineral phases are often decisive for distinguishing earth pigments of similar colour but different geological origin. However, bulk spectral approaches implicitly assume mineral homogeneity and may underrepresent minor components that are critical for discrimination. By combining particle-resolved Raman spectroscopy with statistically robust sampling (>200 spectra per sample), we generate normalized mineral occurrence profiles that capture compositional variability at the particle level rather than relying on a single, averaged spectrum.

Systematic benchmarking of three post-acquisition identification strategies, manual visual identification, automated three-component spectral library fitting (3C model), and supervised machine learning classification (XGBoost model), showed that consistent mineral fingerprint detection is essential for reliable pigment discrimination. Visual identification, while flexible, proved insufficient to separate all five pigments via multi-mineral K-means cluster analysis. By contrast, both semi-automated approaches (3C and XGBoost) produced reproducible mineral assignments, enabling statistical differentiation of all five pigments. The 3C model may inflate minor components when spectral optimization begins to model noise, whereas XGBoost performance depends on representative training data, underscoring that each approach embodies distinct assumptions and trade-offs.

Although particle-specific Raman acquisition increases measurement time compared with bulk analysis, it enables statistically meaningful discrimination of accessory mineral content using a single, non-destructive technique. This approach enables rapid preliminary

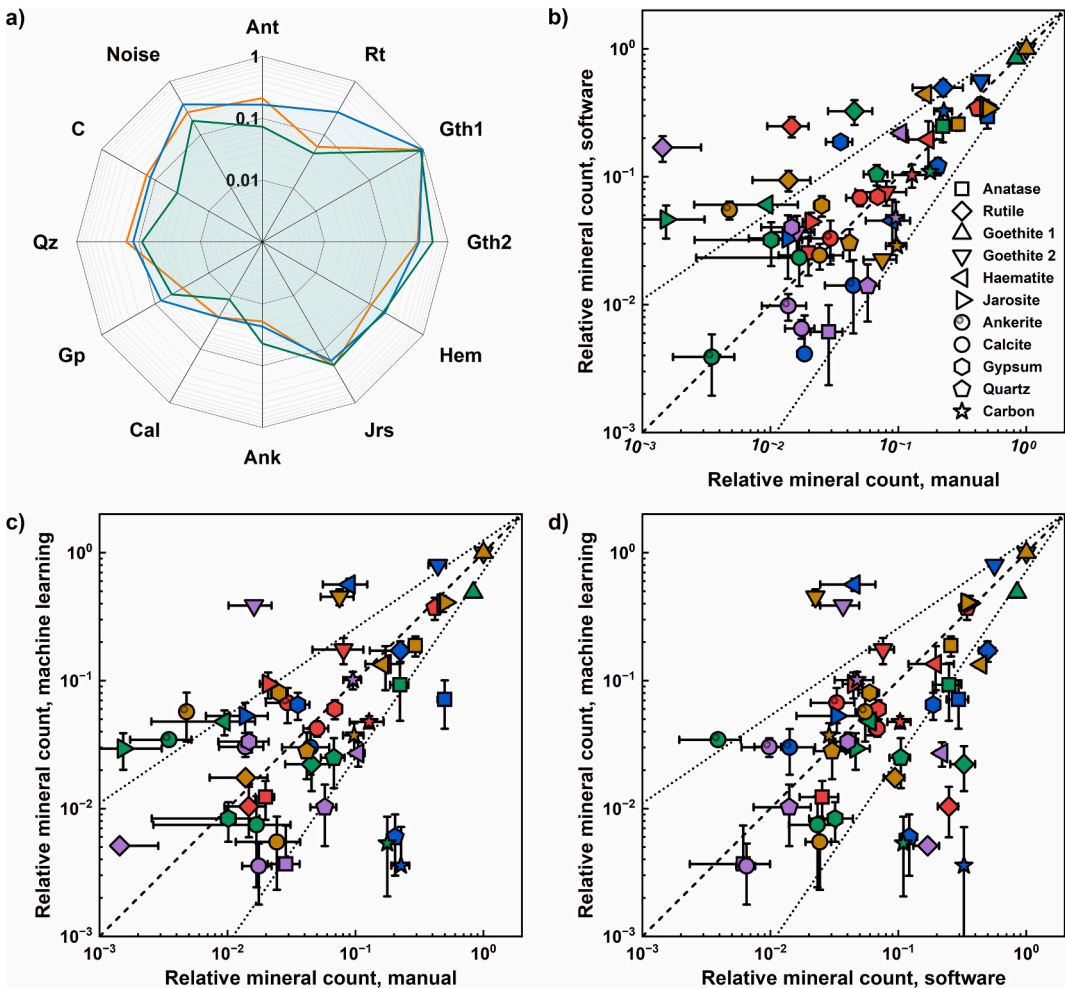


Fig. 5. Comparison of the three analytical methods: **a)** Radar plot of the average relative mineral content for each analytical method (orange: manual; blue: 3C; green: XGBoost); Comparison of relative mineral content, subdivided by mineral (see key) and pigment (red: YOA; blue: YOB; green: YOG; purple: RSD; yellow: RSI), as determined by **b)** 3C vs visual identification, **c)** XGBoost vs visual identification, and **d)** XGBoost vs 3C. $N = 3$, $n \geq 200$, error bars: S.E, dashed line: $y = x$, dotted lines: plus/minus one additional count per 100 counts of the dominant mineral. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Advantages and disadvantages of each analytical method.

Methodology	Advantages	Disadvantages
Visual identification	• Risk of overfitting is low • New mineral spectra can be identified by consulting literature during analysis • Background subtraction is not required for component identification	• Time intensive • Highly dependent on expertise of user • Risk of unconscious bias in mineral assignment • Challenging to identify overlapping components and components with low signal
3C model	• Rapid processing compared to visual identification • Minimal user expertise required • Mathematical model enables accurate deconvolution of overlapping bands	• Complete database of relevant spectra required • Limitations in maximum number of materials assigned (software specific) • Calculation of mathematical fits is computationally expensive if large databases are used • Software may assign negative spectra, which is not experimentally feasible • Risk of overfitting due to optimization based on minimization of difference between the model and the data • Background subtraction required, which may introduce uncertainty into the data
XGBoost model	• Rapid processing compared to visual identification • Minimal user expertise required once model is trained	• Complete database of relevant spectra required • Initial training of model computationally expensive • Training must consider all possible spectral scenarios to ensure accuracy • Background subtraction required, which may introduce uncertainty into the data

pigment differentiation, recognizing that additional analytical techniques, including FTIR, LIBS, SEM-EDS and diffraction methods, are still required to fully characterize each pigment, especially with respect to the identification of aluminosilicates. Overall, this study supports the

use of systematic post-acquisition processing of individual particle spectra as a practical route to improve Raman-based mineral identification and pigment differentiation in cultural heritage science.

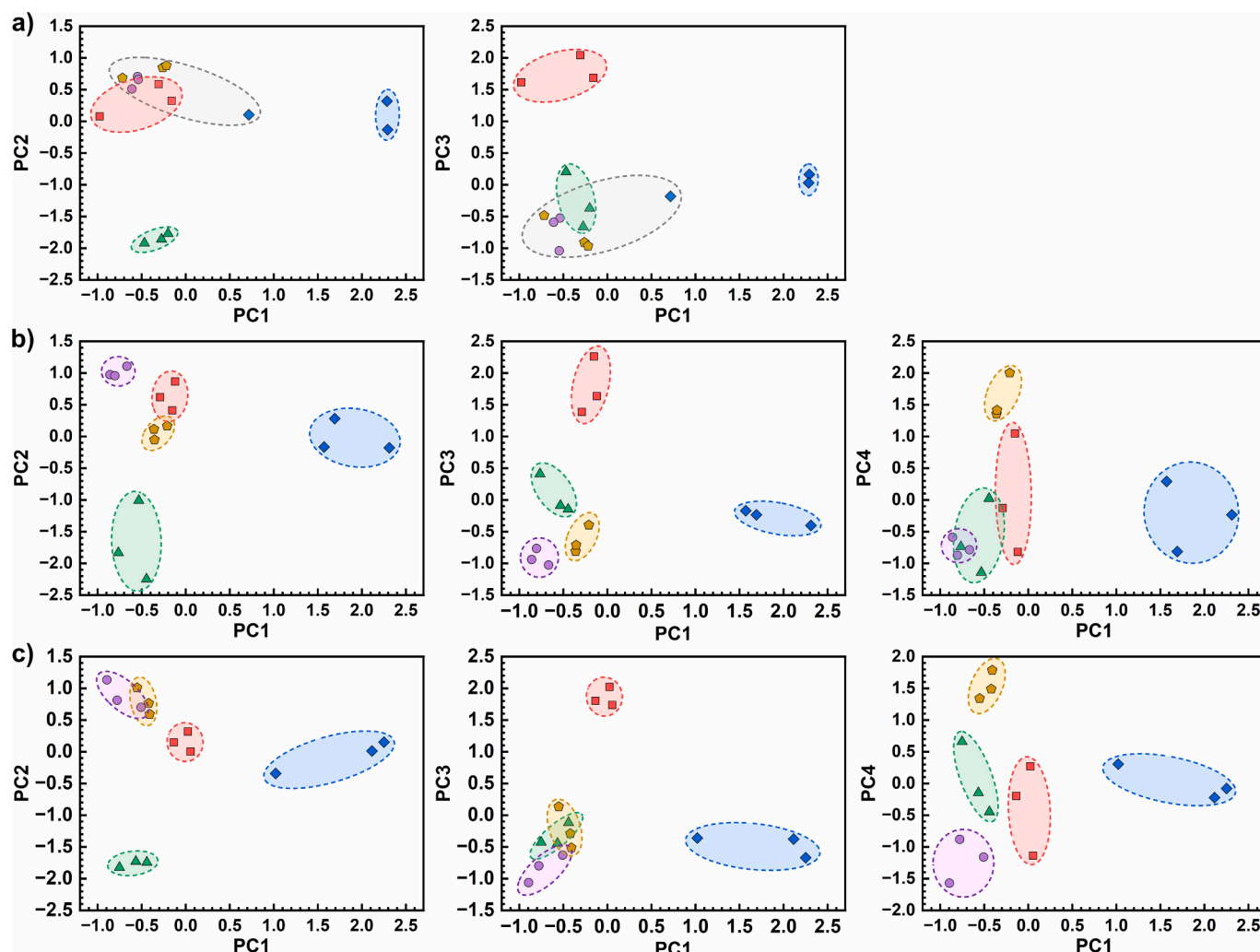


Fig. 6. Principal component (PC) plots for each identification method: **a)** visual identification; **b)** 3C; **c)** XGBoost. Pigment assigned by marker colour and shape (red square: YOA; blue diamond: YOB; green triangle: YOG; purple circle: RSD; yellow pentagon: RSI). Banded areas indicate assigned clusters according to K-means cluster and silhouette analysis. Colored areas indicate clusters that only contain a single pigment, whilst grey areas contain more than one pigment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgements

This research was financially supported by the Interreg 6a Deutschland-Denmark Program project TORCH (grant number 07-1-221). Raman spectroscopy was supported by the national infrastructure UFM 5229-00010B, NANOCHM, and multiple Danfoss Foundation for Education grants. FH and CK also acknowledge support from the Region of Southern Denmark ("Den Syddanske Udviklingspulje") project AutoSERS.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.saa.2026.128238>.

Data availability

Data will be made available on request.

References

- [1] G.P. Mastrotheodoros, K.G. Beltsios, Pigments—Iron-based red, yellow, and brown ochres, *Archaeol. Anthropol. Sci.* 14 (2022) 35.
- [2] R. Siddall, Mineral pigments in archaeology: their analysis and the range of available materials, *Minerals* 8 (2018) 201.
- [3] M. Elias, C. Chartier, G. Prévot, H. Garay, C. Vignaud, The colour of ochres explained by their composition, *Mater. Sci. Eng. B* 127 (2006) 70–80.
- [4] F. Froment, A. Tournié, P. Colomban, Raman identification of natural red to yellow pigments: ochre and iron-containing ores, *J. Raman Spectrosc.* 39 (2008) 560–568.
- [5] G. Vita, M.L. Saladino, F. Armetta, L. Sineo, Geochemical and mineralogical characterization on an ochre residue adhering to a pebble found in the Oriente A Epigravettian burial, in the Grotta D'oriente of Favignana (Egadi, Italy), *Archaeol. Anthropol. Sci.* 16 (2024) 175.
- [6] S.E. Zarzycka, T.A. Surovell, M.E. Mackie, S.R. Pelton, R.L. Kelly, P. Goldberg, J. Dewey, M. Kent, Long-distance transport of red ochre by Clovis foragers, *J. Archaeol. Sci. Rep.* 25 (2019) 519–529.
- [7] L. Burgio, Pigments, dyes and inks: their analysis on manuscripts, scrolls and papyri, *Archaeol. Anthropol. Sci.* 13 (2021) 194.
- [8] G. Chiari, D. Scott, Pigment analysis: potentialities and problems, *Period. Mineral.* 73 (2004) 227–237.
- [9] M.L. Franquelo, A. Duran, L.K. Herrera, M.C. Jimenez de Haro, J.L. Perez-Rodriguez, Comparison between micro-Raman and micro-FTIR spectroscopy

- techniques for the characterization of pigments from southern Spain cultural heritage, *J. Mol. Struct.* 924-926 (2009) 404–412.
- [10] I. Żmuda-Trzebiatowska, M. Wachowiak, A. Klisińska-Kopacz, G. Trykowski, G. Śliwiński, Raman spectroscopic signatures of the yellow and ochre paints from artist palette of J. Matejko (1838–1893), *Spectrochim. Acta A* 136 (2015) 793–801.
 - [11] J. Riu, B. Giussani, Analytical chemistry meets art: the transformative role of chemometrics in cultural heritage preservation, *Chemom. Intel. Lab. Syst.* 247 (2024) 105095.
 - [12] M.C. Caggiani, M. Fugazzotto, L. de Ferri, A. Bertino, F. Andriulo, G. Barone, P. Mazzoleni, Exploitation of IR statistical results as a guidance for the multi-technique study of synthetic and natural iron-based pigments, *Microchem. J.* 197 (2024) 109851.
 - [13] I.M. Cortea, L. Ghervase, R. Rădvan, G. Serîţan, Assessment of easily accessible spectroscopic techniques coupled with multivariate analysis for the qualitative characterization and differentiation of earth pigments of various provenance, *Minerals* 12 (2022) 755.
 - [14] M.K. Donais, D. George, B. Duncan, S.M. Wojtas, A.M. Daigle, Evaluation of data processing and analysis approaches for fresco pigment studies by portable X-ray fluorescence spectrometry and portable Raman spectroscopy, *Anal. Methods* 3 (2011) 1061–1071.
 - [15] P.M. Ramos, I. Ruisánchez, K.S. Andrikopoulos, Micro-Raman and X-ray fluorescence spectroscopy data fusion for the classification of ochre pigments, *Talanta* 75 (2008) 926–936.
 - [16] P. Jahoda, I. Drozdovskiy, S.J. Payler, L. Turchi, L. Bessone, F. Sauro, Machine learning for recognizing minerals from multispectral data, *Analyst* 146 (2021) 184–195.
 - [17] M. Veneranda, J.A. Manrique, A. Sanz-Arranz, S. Julve Gonzalez, C. Prieto Garcia, E. Pascual Sanchez, M. Konstantinidis, E. Charro, J.M. Lopez, M.A. Gonzalez, F. Rull, G. Lopez-Reyes, Application of chemometrics on Raman spectra from Mars: recent advances and future perspectives, *J. Chemometr.* 37 (2023) e3438.
 - [18] I.M. Bell, R.J.H. Clark, P.J. Gibbs, Raman spectroscopic library of natural and synthetic pigments (pre- ≈ 1850 AD), *Spectrochim. Acta A* 53 (1997) 2159–2179.
 - [19] D. Bikiaris, S. Daniilia, S. Sotiropoulou, O. Katsimbiri, E. Pavlidou, A. P. Moutsatsou, Y. Chrysoulakis, Ochre-differentiation through micro-Raman and micro-FTIR spectroscopies: application on wall paintings at Meteora and Mount Athos, Greece, *Spectrochim. Acta A* 56 (2000) 3–18.
 - [20] M. Bouchard, D.C. Smith, Catalogue of 45 reference Raman spectra of minerals concerning research in art history or archaeology, especially on corroded metals and coloured glass, *Spectrochim. Acta A* 59 (2003) 2247–2266.
 - [21] L. Burgo, R.J.H. Clark, Library of FT-Raman spectra of pigments, minerals, pigment media and varnishes, and supplement to existing library of Raman spectra of pigments with visible excitation, *Spectrochim. Acta A* 57 (2001) 1491–1521.
 - [22] M.C. Caggiani, A. Cosentino, A. Mangone, Pigments checker version 3.0, a handy set for conservation scientists: A free online Raman spectra database, *Microchem. J.* 129 (2016) 123–132.
 - [23] K. Castro, M. Pérez-Alonso, M.D. Rodríguez-Laso, L.A. Fernández, J.M. Madariaga, On-line FT-Raman and dispersive Raman spectra database of artists' materials (e-VISART database), *Anal. Bioanal. Chem.* 382 (2005) 248–258.
 - [24] A. Coccato, M.C. Caggiani, An overview of principal components analysis approaches in Raman studies of cultural heritage materials, *J Raman Spectrosc* 55 (2024) 125–147.
 - [25] B. Lafuente, R.T. Downs, H. Yang, N. Stone, The Power of Databases: The RRUFF Project, in: T. Armbruster, R.M. Danisi (Eds.), *Highlights in Mineralogical Crystallography*, W. De Gruyter, Berlin, Germany, 2015, pp. 1–30.
 - [26] G. Marucci, A. Beeby, A.W. Parker, C.E. Nicholson, Raman spectroscopic library of medieval pigments collected with five different wavelengths for investigation of illuminated manuscripts, *Anal. Methods* 10 (2018) 1219–1236.
 - [27] L.I. McCann, K. Trentelman, T. Possley, B. Golding, Corrosion of ancient Chinese bronze money trees studied by Raman microscopy, *J Raman Spectrosc* 30 (1999) 121–132.
 - [28] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, M.D. Dyar, Machine learning tools for mineral recognition and classification from Raman spectroscopy, *J. Raman Spectrosc.* 46 (2015) 894–903.
 - [29] W. Dong, M. Qin, S. Wu, L. Hu, C. Rao, Z. Du, Deep learning-assisted Raman spectroscopy for automated identification of specific minerals, *Spectrochim. Acta A* 333 (2025) 125843.
 - [30] J.J. González-Vidal, R. Pérez-Pueyo, M.J. Soneira, S. Ruiz-Moreno, Independent component analysis-based algorithm for automatic identification of Raman spectra applied to artistic pigments and pigment mixtures, *Appl. Spectrosc.* 69 (2015) 314–322.
 - [31] M. Wojdyr, Fityk: a general-purpose peak fitting program, *J. Appl. Cryst.* 43 (2010) 1126–1128.
 - [32] T. Mi, J.J. Wang, C. McCague, Y. Bai, Application of Raman spectroscopy in the study of the corrosion of steel reinforcement in concrete: A critical review, *Cem. Concr. Compos.* 143 (2023) 105231.
 - [33] F. Wu, F. Liu, J. Zhou, F. Meng, C. Liu, Structural incorporation of aluminum into goethite and its impact on selenium adsorption, *Chem. Eng. J.* 505 (2025) 159517.
 - [34] D. Li, B.H. O'connor, I.-M. Low, A.V. Riessen, B.H. Toby, Mineralogy of Al-substituted goethites, *Powder Diffract* 21 (2006) 289–299.
 - [35] A. Burkov, Machine Learning Engineering: True Positive Incorporated, 2020.