Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2023, **148**, 2861

Non-destructive distinction between geogenic and anthropogenic calcite by Raman spectroscopy combined with machine learning workflow†

Sara Calandra, *[a,b] Claudia Conti, [c] Irene Centauro [a] and Emma Cantisani [d]

Here, we demonstrate, for the first time, the possibility of distinguishing between geogenic and anthropogenic calcite in a non-destructive and effective way. Geogenic calcite derives from natural sedimentary and metamorphic rocks whereas anthropogenic calcite is formed artificially due to the carbonation process in mortars and plaster lime binders. Currently, their distinction is a major unaddressed issue although it is crucial across several fields such as ^{14}C dating of historical mortars to avoid contamination with carbonate aggregates, investigating the origins of pigments, and studying the origins of sediments, to name a few. In this paper, we address this unmet need combining high-resolution micro-Raman spectroscopy with data mining and machine learning methods. This approach provides an effective means of obtaining robust and representative Raman datasets from which samples' origins can be effectively deduced; moreover, a distinction between sedimentary and metamorphic calcite has been also highlighted. The samples, chemically identical, exhibit systematic and reliable differences in Raman band positions, band shape and intensity, which are likely related to the degree of structural order and polarization effects.

Received 21st March 2023,
Accepted 12th May 2023

DOI: 10.1039/d3an00441d

rsc.li/analyst

Introduction

Calcite is a mineral widely diffused on the Earth's surface, having different origins. It is mainly present in sedimentary and metamorphic rocks (*e.g.* marbles) and can also be produced by biological systems and human activities (pyrotechnology origin).

Anthropogenic calcite is mainly found as a binder in mortars and plasters and is produced following traditional technologies.[1,2] The production of lime mortar is shown in reactions (1)–(3). Air-hardening calcitic limes are obtained by: (1) burning pure limestones at temperatures of 800–950 °C; (2) hydration of calcium oxide; and then, (3) carbonation of calcium hydroxide in air with the formation of calcite.

$$CaCO_3 + \text{heat} = CaO\ (\text{lime}) + CO_2 \quad (1)$$

$$CaO + H_2O = Ca(OH)_2\ (\text{portlandite}) + \text{heat} \quad (2)$$

$$Ca(OH)_2 + CO_2 = CaCO_3 + H_2O. \quad (3)$$

The calcite obtained in this process has the same chemical composition as burnt limestone, but has different textural and mechanical properties. An efficient, fast, effective and widespread technique in laboratories is needed to distinguish calcite from different domains.

Satisfactory results were obtained using Fourier transform infrared spectroscopy (FTIR) in different configurations[3–5] and the luminescence properties of calcium carbonate.[6,7] Both methods are based on different densities and distributions of atomic defects in the calcite crystal structure.

FTIR can distinguish calcite formed by different processes using the trend lines of anthropogenic and geogenic calcite constructed from the intensity of specific bands.[3,4] Luminescence allows us to identify the structural defects in calcite that cause changes in the infrared spectra. The ion substitutions provide luminescence activators or quenchers. Most geogenic forms of $CaCO_3$, *e.g.* limestone, exhibit red–orange luminescence due to the presence of Mn^{2+} sites in the calcite

[a]*Department of Chemistry Ugo Schiff, University of Florence, 50019 Sesto Fiorentino, Italy. E-mail: sara.calandra@unifi.it*
[b]*Department of Earth Sciences, University of Florence, 50121 Florence, Italy. E-mail: sara.calandra@unifi.it, irene.centauro@unifi.it*
[c]*Institute of Heritage Science (ISPC), National Research Council (CNR), 20125 Milano, Italy. E-mail: claudia.conti@cnr.it*
[d]*Institute of Heritage Science (ISPC), National Research Council (CNR), 50019 Sesto Fiorentino, Italy. E-mail: emma.cantisani@cnr.it*

† Electronic supplementary information (ESI) available: SFig. 1: Figure depicting the 2D plots of v_1, v_4, L wavenumbers are reported and expressed as average values; SFig. 2: Scatterplot from the key influence factor visual: increase in % of anthropogenic calcite samples of v_4 intensity values. For more details about data analysis, see DOI: https://doi.org/10.1039/d3an00441d

crystal lattice.[8] In anthropogenic calcite, the luminescence varies, since its formation process involves molecular structure changes, decreasing the number of luminescence centers in the structure.[9,10]

This study aims to verify the feasibility of micro-Raman spectroscopy to distinguish calcite formed by different processes. The Raman spectrum of calcite is characterized by a ν_1 sharp band at 1086 cm^{-1} along with other subsidiary bands at 156 cm^{-1} (T), 282 cm^{-1} (L) and 712 cm^{-1} (ν_4).[11–17]

The Raman technique was used to estimate the cation (Mg^{2+}, Fe^{2+}, and Mn^{2+}) content in carbonate since the vibrational wavenumbers of the translational (T) and librational (L) modes of carbonates are sensitively related to their cation composition,[13,17,18] to investigate the changes in atomic bonds in biogenic calcite crystals,[19] and to distinguish the degree of crystallinity of calcium carbonate in biological materials,[20] evaluating the wavenumbers and the width of v_1 and v_4 bands.

These papers highlight the suitability of Raman spectroscopy for evidencing the structural and chemical changes that occur in the calcite lattice. Indeed, by studying the variation of the structure of calcite, the short-range order is best detected at the molecular level using Raman spectroscopy.[21]

The micro-Raman identification of anthropogenic calcite can be used for different purposes: (1) for the selection of the datable fraction from binders in aerial mortars, avoiding any type of contamination with geogenic calcite due to the presence of carbonate aggregates or the remains of underburnt fragments of stone for lime – the accurate ^{14}C dating of mortars is strictly related to the removal of this kind of contaminant;[10,22,23] (2) to distinguish the preparation technique of white pigments (crushed rocks or lime-based materials); (3) to identify calcitic wood ash in sediments;[3] and (4) to identify the self-healing areas in ancient mortars.[24] In these frameworks, since a very small amount of material samples is available, which must be preserved for further analyses, a non-destructive high-resolution micro-Raman technique is recommended and strongly encouraged. In addition, the Raman technique in the portable configuration is more easily applicable than the respective IR spectroscopy (diffuse reflectance spectroscopy) to broaden the use of the calcite identification method in a non-invasive way. It is known that portable FTIR provides spectral modifications, such as distortion, inversion, enhancement, or abatement of infrared bands,[25] which can hinder our application.

We selected a wide range of different geogenic and anthropogenic calcites, belonging to different carbonate rocks and mortars. We used high-resolution micro-Raman spectroscopy to accurately measure the order of crystal calcite and identify the information on the spectrum of the geogenic calcite and anthropogenic calcite.

Two technologies, Microsoft Power BI and Python, were used to build a data analysis workflow aiming to distinguish groups of the spectral data acquired for the different calcite samples and to identify their characteristic Raman spectral features. Another objective of the data analysis was to evaluate the accuracy of the identification of geogenic and anthropogenic calcite from spectral data through a comparison between machine learning models.

Materials and methods

Selected samples

The selected samples consist of calcite belonging to Italian geological materials (geogenic calcite samples) and calcite extracted from the binders of air-hardening mortar samples (anthropogenic calcite samples). 13 carbonate rocks, generally burnt to produce quicklime, taken from different Italian quarries and 11 binder mortars collected from historical buildings, factory-made binders and test specimens made in the laboratory were investigated (Table 1). Lumps represent portions of unmixed lime in an aerial mortar produced with traditional technologies.[2]

In order to systematically investigate the powder, the particle sizes were controlled through different sieves, up to a granulometric class below 25 μm.[26]

All samples were first analysed through X-ray powder diffractometry (XRPD), scanning electron microscopy with energy dispersive X-ray spectroscopy (SEM-EDS) and attenuated total reflection Fourier transform infrared spectroscopy (ATR-FTIR) and subsequently beamed under a Raman spectrometer. This extensive investigation, not reported here, was performed to control the composition of powders for the suitable selection of those samples consisting mainly of calcite. In fact, the reduction of the number of variables, and the consequent complexity of the system, is essential at this stage for the proper interpretation of spectral changes.

In addition, lime binders and lumps extracted from ancient mortars were thoroughly characterized by optical microscopy (OM), thermogravimetric analysis (TGA) and optical microscopy–cathodoluminescence (OM-CL) imaging to evaluate their reliability for this study (results are not reported here).

Raman spectroscopy

Raman spectra were collected using a high-resolution Renishaw inVia Raman spectrometer coupled to a Leica DMLM microscope. The measurements were carried out with a 785 nm excitation line equipped with a 50× long working distance objective (NA 0.5, a spectral resolution of <1 cm^{-1} and a theoretical laser spot diameter of 1.9 μm). A laser power of 80 mW and an acquisition time of 5 s per spectrum were used.

We decided to focus our attention towards the low-medium region of the spectral range, collected in the range of 100–1400 cm^{-1}. For each powder, we took 10 Raman spectra at slightly different positions.

The wavenumbers, intensities, and areas of typical vibrations of carbonate groups in calcite (L, librational mode; v_4, in-plane bending mode; and v_1, symmetric stretching mode) were processed with Spectragryph v 1.2.15 software.

Table 1 List samples, reporting the ID sample, material type and provenance, sample composition and calcite type (geogenic or anthropogenic)

ID sample	Material type and provenance	Composition[a]	Calcite type[b]
MAR	Marble, Carrara (Tuscany, Italy)	Cal (+++)	Geogenic
CAMP 1	Marble, Campiglia Marittima (Tuscany, Italy)	Cal (+++)	Geogenic
CAMP 2	Marble, Campiglia Marittima (Tuscany, Italy)	Cal (+++)	Geogenic
CAMP 3	Marble, Campiglia Marittima (Tuscany, Italy)	Cal (+++)	Geogenic
MS	Marble, Montagnola Senese (Tuscany, Italy)	Cal (+++)	Geogenic
LIM	Marble, Carrara (Tuscany, Italy)	Cal (+++), qz (*)	Geogenic
PLEC	Limestone, *Pietra di Lecce* (Apulia, Italy)	Cal (+++)	Geogenic
ALB L	Limestone, *Alberese*, Monte Morello (Tuscany, Italy)	Cal (+++), cl min (*), qz (*)	Geogenic
ALB A	Limestone, *Alberese*, Monte Morello (Tuscany, Italy)	Cal (+++), cl min (*), qz (*)	Geogenic
TRAV	Travertine, Rapolano (Tuscany, Italy)	Cal (+++), qz (*)	Geogenic
PGAL	Limestone, *Pietra Gallina* (Venetian region, Italy)	Cal (+++)	Geogenic
PMAT	Limestone, *Pietra di Matera* (Basilicata, Italy)	Cal (+++)	Geogenic
PVIC	Limestone, *Pietra di Vicenza* (Venetian Region, Italy)	Cal (+++)	Geogenic
OS	Ancient plaster, archaeological site	Cal (+++)	Anthropogenic
LS01	Laboratory mortar	Cal (+++), qz (+), portl (*)	Anthropogenic
WHL	Factory-made binder	Cal (+++), cl min (*), qz (*)	Anthropogenic
CT26L1	Lime lump, historical building	Cal (+++), qz (++)	Anthropogenic
CT26L2	Lime lump, historical building	Cal (+++), qz (*)	Anthropogenic
CT26L4	Lime lump, historical building	Cal (+++), qz (*)	Anthropogenic
CT27L4	Lime lump, historical building	Cal (+++)	Anthropogenic
CT27L1	Lime lump, historical building	Cal (+++)	Anthropogenic
SFC1B1	Lime binder, historical church	Cal (+++), qz (*)	Anthropogenic
SFC1L1	Lime lump, historical church	Cal (+++), qz (*)	Anthropogenic
SFC5B1	Lime binder, historical church	Cal (+++), qz (*)	Anthropogenic

Cal: calcite; qz: quartz; cl min: clay minerals; portl: portlandite. +++: very abundant; ++: abundant; +: present; *: traces; and –: below the detection limit. [a] *Via* XRPD, SEM-EDS and TGA. [b] *Via* OM, OM-CL, and ATR-FTIR.

The spectra were not baseline corrected but normalized to the v_1 height. For each Raman spectrum, from the L, v_4, and v_1 bands were collected (Fig. 1): (i) the position of the band, to evaluate the wavenumber shift; (ii) the intensity of band, following the method of Chu *et al.* (2008),[3] where the intensity value was subtracted from the specific baseline; and (iii) the area subtended by the band without the baseline.

Given the amount of variables, the extracted parameters were used for statistical analysis of the data to investigate the presence of discriminating factors for distinguishing geogenic and anthropogenic calcite. To better investigate the obtained results, full-width at half-maximums (FWHMs) were recorded for L, v_4 and ν_1 Raman bands for each sample.

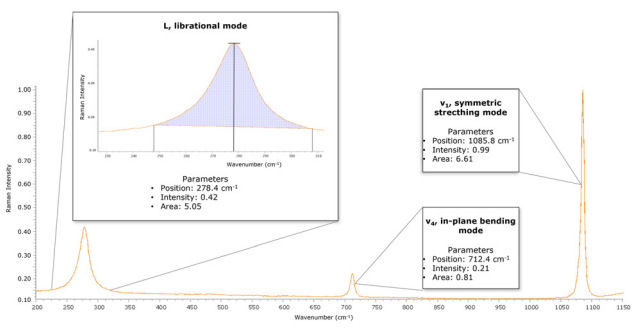

Fig. 1 A selected region of measured micro-Raman spectra of the carbonate samples. Wavenumbers, intensities, and areas of L, v_4, v_1 were collected as shown in the detailed windows at 230 and 310 cm^{-1}, which highlight the data-taking method.

Data exploration and analysis

The proposed workflow integrates Microsoft Power BI data visualization and analysis tool and Python programming language with the Scikit-learn package.[27,28] The proposed method involves the following main steps: (1) visual inspection of the dataset; (2) reduction of the dataset dimensionality and segmentation by principal component analysis (PCA) and K-means clustering; and (3) building of machine learning models able to predict the value of the target variable (calcite types) based on the values of the independent variables (logistic regression and random forest models).

Raman spectra data are stored in a dataframe: each parameter collected from Raman spectra is called a "feature" (or a "variable"); the 2 possible classes of the target variable are geogenic or anthropogenic calcite. For each variable, outliers are detected and removed by the interquartile range (IQR) method, calculated in Python.[29]

Then, visual inspection is carried out in Power BI, directly connected to the dataframe, through the key influence factor (KIF) visual, which performs ML.NET SDCA regression implementation.[30] According to the second step, PCA was performed in Python, using the sklearn.decomposition.PCA function.[31]

Before applying the PCA, data are standardized using StandardScaler, a function implemented in the Scikit-learn package, so that all features are at the same scale. From the transformed dataframe after PCA, K-means clustering in Power BI clustering visual is performed. Then, the dataframe is randomly divided into a training set and a testing set (with a 70 : 30 split ratio) in Python. A comparison between logistic

regression and random forest models[31–34] is performed in Python, with the Scikit-learn functions logistic regression and random forest classifier, on the PCA components, setting up a repeated K-fold cross-validation with the Scikit-learn function K-fold on the training set, in order to find the best fit to describe the relationship between the target variable and the predictor variables.

Results and discussion

Analytical characterization of the geogenic and anthropogenic calcites

The Raman calcite spectrum (Fig. 1 and 2a) is characterized by an intense band at 1086 cm^{-1} (v_1), along with other subsidiary bands: a weak band at 712 cm^{-1} (v_4) and a medium intensity band at 282 cm^{-1} (L). These wavenumbers are characteristic of the samples consisting of geogenic calcite, and are not observed in anthropogenic calcite samples, as they exhibit a Raman shift at bands L and v_1. Geogenic samples present on average an L varying from 280.4 to 282.4 cm^{-1}, a v_4 varying from 712.4 to 713.0 cm^{-1}, and finally a v_1 varying from 1086.2 to 1086.9 cm^{-1} (Table 2). Meanwhile, the anthropogenic samples have an average L ranging from 273.8 to 278.3 cm^{-1}, a v_4 rather constant from 712.1 to 712.5 cm^{-1}, and finally a v_1 ranging from 1085.4 to 1086.0 cm^{-1}. In Fig. 2b, the marble (blue spectrum) is compared with one of the ten spectra obtained from each anthropogenic sample studied, and a significant variation is highlighted, especially, for L and v_1 wavenumbers.

This systematic discrepancy observed in the Raman shifts of the two calcite groups of different origin prompted us to further investigate the information gathered from the main vibrational modes (in Fig. S1,† 2D plots of the main discriminating parameters are reported and expressed as average values). We determined the wavenumber, intensity, and area of the three main vibrational modes of 24 calcite samples. In Table 2, the parameter average is collected by the spectra. A preliminary observation of results suggests differences between the data gathered.

Data analysis results

In the first step of the data analysis workflow, visual inspection of the dataset is performed through the key influence factor (KIF) visual. The KIF highlights the L wavenumber and v_1 wavenumber as the most important influencers to discriminate geogenic from anthropogenic calcite (Fig. 3). The scatterplot in Fig. 3a shows that all the samples with an L wavenumber value over about 280.0 cm^{-1} are of geogenic calcite. Similarly, Fig. 3b shows that 85% of samples with a v_1 wavenumber value higher than 1086.2 cm^{-1} consist of geogenic calcites. In addition, another influencing factor could be the v_4 intensity: samples with v_4 intensity values between 0.026 and 0.098 consist more of anthropogenic calcite than geogenic calcite (Fig. S2†). However, this value is quite variable in geogenic calcite samples, so some fall into this range.

From the preliminary KIF results, correlations between the L and v_1 wavenumbers and v_4 intensity are evaluated. Bubble charts are used to determine whether there is a correlation or a shared trend between at least 3 variables. The L and v_1 wavenumbers seem to be the most significant parameters in discriminating geogenic from anthropogenic calcite, thus they are set as the x and y axes in the bubble chart visual in Power BI (Fig. 4).

A different distribution of samples is clearly visible in Fig. 4a, where geogenic samples (in blue) are all located over about 280.0 cm^{-1} (L wavenumber) and 1086.2 cm^{-1} (v_1 wave-

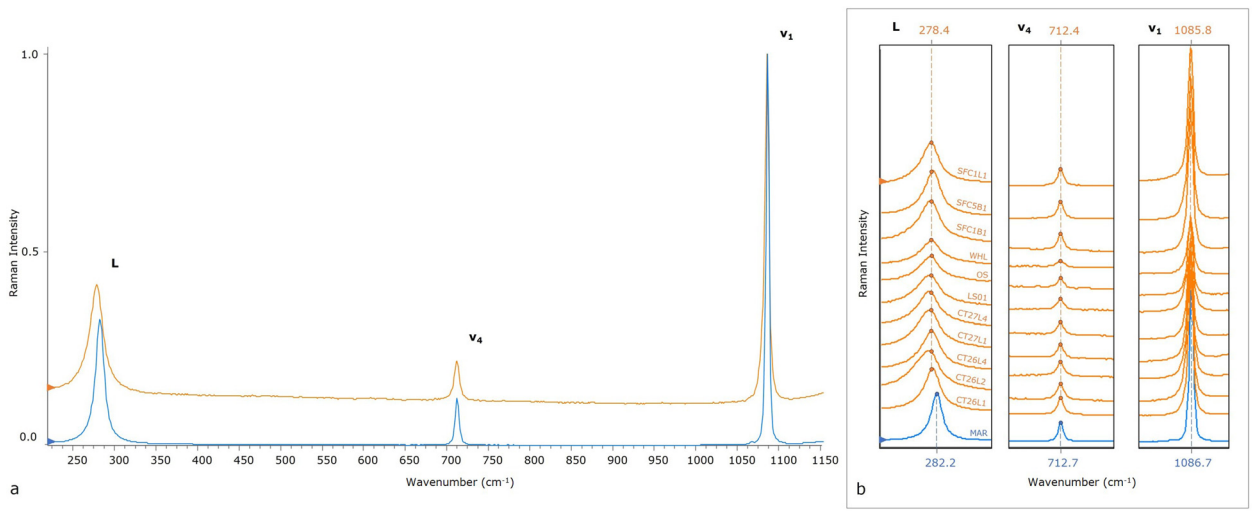

Fig. 2 Comparison among individual Raman spectra of carbonate samples, normalized to the v_1 intensity: geogenic calcite (in blue, MAR sample) and anthropogenic calcite (in orange, SFC1L1 sample) (a). Raman spectra details for the anthropogenic sample (in orange) are shown with the geogenic sample (b). Band positions of L, v_4, v_1 are reported to highlight the Raman shift, especially in L vibrational mode. Bottom, characteristic values of the band position of the MAR sample (blue), and top, characteristic values of the SFC1L1 sample (first orange band).

Table 2 The average of the variables: the wavenumbers, intensities, and areas of typical vibrations of carbonate groups in calcite were collected from 10 Raman measures performed for each sample

ID sample	L_wavenumber	L_intensity	L_area	v_4_wavenumber	v_4_intensity	v_4_area	v_1_wavenumber	v_1_intensity	v_1_area
MAR	282.4	0.29	4.07	712.9	0.10	0.68	1086.8	0.94	4.81
CAMP 1	281.9	0.29	4.38	712.5	0.09	0.54	1086.4	0.86	4.24
CAMP 2	280.9	0.27	4.44	712.5	0.10	0.62	1086.5	0.91	5.01
CAMP 3	281.5	0.23	3.48	712.4	0.08	0.49	1086.4	0.75	3.73
MS	282.1	0.28	4.26	712.6	0.11	0.67	1086.6	0.61	4.88
LIM	280.4	0.28	5.56	712.7	0.11	0.84	1086.4	0.96	6.71
PLEC	281.9	0.06	1.51	713.0	0.01	0.08	1086.9	0.14	0.68
ALB L	281.4	0.05	1.53	712.8	0.01	0.09	1086.7	0.11	0.57
ALB A	281.1	0.06	1.72	712.5	0.01	0.08	1086.5	0.19	0.78
TRAV	281.9	0.19	3.20	712.8	0.07	0.43	1086.7	0.60	3.44
PGAL	282.0	0.09	1.73	712.7	0.02	0.14	1086.7	0.24	1.24
PMAT	282.0	0.09	1.80	712.5	0.03	0.18	1086.6	0.28	1.47
PVIC	281.4	0.13	2.28	712.4	0.05	0.27	1086.2	0.41	2.08
OS	276.4	0.08	2.41	712.2	0.02	0.19	1085.6	0.20	1.59
LS01	277.6	0.09	2.11	712.3	0.03	0.30	1085.8	0.36	2.37
WHL	276.4	0.18	5.14	712.3	0.06	0.52	1085.6	0.66	5.15
CT26L1	275.0	0.14	4.02	712.1	0.05	0.42	1085.4	0.46	3.36
CT26L2	277.4	0.21	5.34	712.4	0.08	0.58	1085.8	0.68	4.14
CT26L4	277.2	0.21	5.08	712.4	0.08	0.55	1085.8	0.77	5.04
CT27L4	273.8	0.20	6.46	712.2	0.07	0.62	1085.4	0.62	5.11
CT27L1	277.8	0.25	5.99	712.5	0.09	0.70	1086.0	0.87	5.76
SFC1B1	277.5	0.18	4.67	712.5	0.06	0.53	1085.9	0.62	4.89
SFC1L1	277.7	0.23	5.65	712.4	0.08	0.66	1085.8	0.77	5.92
SFC5B1	278.3	0.24	5.42	712.5	0.09	0.69	1085.9	0.78	5.73

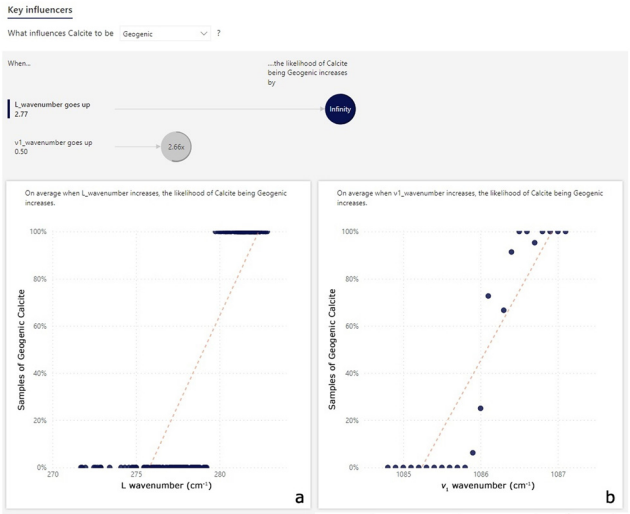

Fig. 3 Scatterplots and a list of top influencers from the key influence factor visual of geogenic calcite samples. The upper part of the figure shows that: L wavenumber is the top factor contributing to identification of the geogenic calcite samples (*i.e.* the likelihood of calcite being geogenic increases by infinite times). This trend is best explained by the scatterplot in (a). The second factor is the v1 wavenumber (*i.e.* the samples are 2.66 times more likely to consist of geogenic calcite when the value increases), as highlighted in the scatterplot in (b).

number). To better highlight the behavior of the v_4 intensity of the KIF results, Fig. 4b is built with only geogenic calcite samples and it can be observed that the samples are well separated along the *x*-axis (v_4 intensity). Most of the samples have v_4 intensity values below and above the range of 0.026–0.098, in which the majority of anthropogenic calcite samples fall, except for the PMAT, PVIC, TRAV, CAMP 3 and CAMP 1 samples. In particular, the v_4 intensity within the geogenic samples is more variable, so this parameter could discriminate against the geogenic calcite types. In addition, the distribution of geogenic samples in the graph could allow a distinction between sedimentary carbonate (ALB A, ALB L, PGAL, PMAT, PLEC, and PVIC) and metamorphic (MAR, CAMP 1, CAMP 2, CAMP 3, LIM, and MS) rocks. Detached from the two groups is the travertine sample (TRAV), since it is a sedimentary rock of chemical origin.

To highlight pairwise relationships between the variables and to complete the visual inspection step in the dataframe, a pairplot is created. Fig. 5 shows that the original dataframe has a high level of multicollinearity, since many variables are strongly correlated with one or more of the other variables. In line with the findings of the KIF analysis, the L and v_1 wavenumbers, correlated with all other parameters, allow us to better distinguish the different calcites. The same consideration can be performed for the v_4 wavenumber, although a precise discriminating factor cannot be considered.

The second step of data analysis provides the PCA in order to eliminate the multicollinearity, reduce the dimensionality of the dataframe and improve the machine-learning algorithm performance. PC1 and PC2 describe 90.2% of the variance (56.7% and 33.5%, respectively). Thus, PCA is performed again keeping only the first 2 PCs. Creating a heatmap of the transformed dataset, it can be seen that no variable is correlated with one or more of the other variables. The Python code is then implemented in Power BI to visualize the biplot of the 2 PCs (Fig. 6). A PCA biplot shows both the PC scores of the samples (dots) and the loadings of the variables (vectors). The

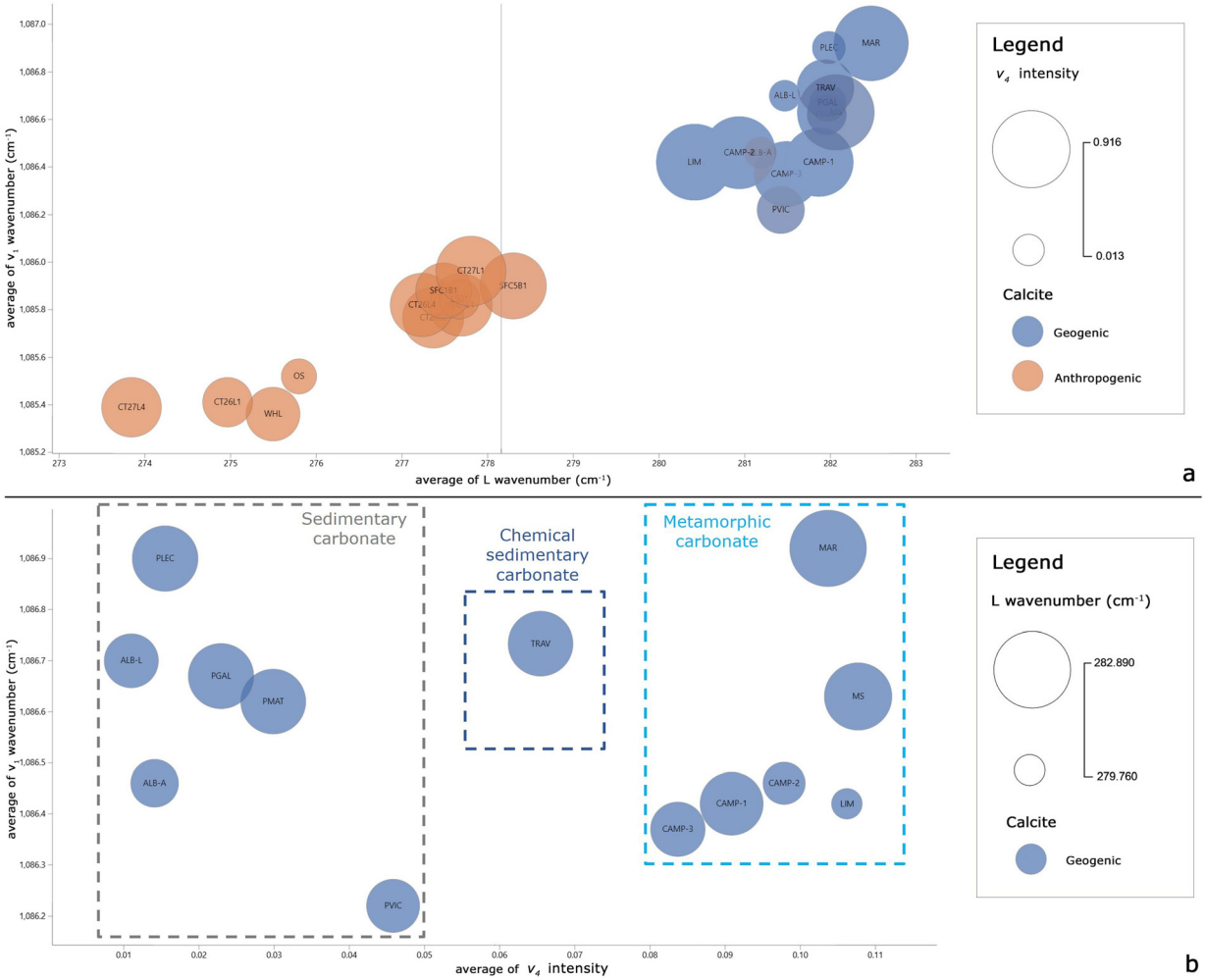

Fig. 4 Bubble charts: average values of v_1 wavenumber *vs.* L wavenumber with v_4 intensities as the bubble size of both calcites (a); average values of v1 wavenumber *vs.* v4 intensities with L wavenumber as the bubble size of geogenic calcites (b).

PC1 *vs.* PC2 scores indicate a clear separation of the geogenic from the anthropogenic calcites, except for a few samples (Fig. 6a). The L, v_1 and v_4 wavenumbers are the most influential variables for calcite distinction and strongly influence the PC2 score.

Another observation on loadings is that the angles between the vectors represent how characteristics correlate with one another: when two vectors are close, forming a small angle, the two variables they represent are positively correlated (*e.g.*, L wavenumber, v_1 wavenumber and v_4 wavenumber), whereas if they are almost perpendicular, they are not likely to be correlated. As already observed in the KIF analysis and bubble charts, the v_4 intensity in the biplot separates geogenic calcites into 3 main groups. From the transformed dataframe after PCA, K-means clustering in Power BI visual is performed (Fig. 6b). K-means is used to identify groups of similar features based on the new representation of data generated by PCA. Of the 5 groups identified, the geogenic calcites are separated into 3 clusters (Fig. 6b), as observed in the bubble chart (Fig. 3b).

PCA and K-means clustering are unsupervised machine learning algorithms that allow us to reduce and segment the data. In order to build a model able to explain the relationship between the target variable (calcite types) and the new variables obtained from PCA, a comparison between supervised machine learning algorithms was performed. Logistic regression classification and random forest classifier algorithms were performed, to extensively investigate the prediction. In general, it is useful to compare different classification or regression models when there are several hypotheses about the relationship between characteristics and the target class and when we want to determine which model provides a better performance for a given classification problem.

The logistic regression algorithm is able to correctly predict 64 out of 67 instances in the test set, resulting in 96% accuracy, while the random forest algorithm is able to correctly predict 62 out of 67 instances, resulting in 93% accuracy (Table 3).

As shown in Table 3, the accuracies of the two models are similar, but logistic regression has higher values of precision,

Fig. 5 Pairplot of the variables. The pairplot is in matrix format where the row name represents the *x* axis and the column name represents the *y* axis; the main-diagonal subplots are the univariate distributions for each attribute.

recall and F1-score, so it seems to be the best model for explaining the relationship between the target variable and the predictor variables, and to predict the binary outcomes. On the other hand, the performance of the random forest model is relatively robust against parameter specifications and less subject to overfitting, because it depends less on parameter values than other machine learning algorithms such as logistic regression.[35]

Discussion

The L and v_1 wavenumbers are the variables which are more influential for distinguishing calcite domains.

These vibrational bands fall into two regions: the L band is due to vibrations of the complete unit cell which are generally referred to as the lattice modes; the v_1 band, is caused by the internal modes of the molecular carbonate ion.[20,36–38]

It is worth noting that Mg is not present in the samples, thus the wavenumber shift is not ascribable to the decrease in the average metal–oxygen bond length (Mg–O bonds are shorter than Ca–O bonds).[13,17,39]

In order to establish more insights, the FWHMs of L, v_4 and v_1 bands were measured, and the average values are reported in Table 4. The carbonate rocks are well-crystallized materials, and the average FWHM for the L band is in the range 11.8–17.4 cm^{-1}; for the v_4 band, it is in the range 5.1–6.8 cm^{-1}; and for the v_1 band, it is in the range 4.3–5.1 cm^{-1}. The binder mortars present significantly higher FWHMs of the L band, in the range 18.1–26.6 cm^{-1}; of v_4 in the range 6.3–8.8 cm^{-1}; and of v_1 in the range 5.2–6.7 cm^{-1}. It is noteworthy that the more the band positions of anthropogenic calcite are shifted to low values, as in the case of L and ν_1 wavenumbers, the higher the FWHM values are.

The relatively large FWHMs reflect the Raman spectroscopic features of a structural disorder in calcite crystals or a small crystalline order: the broader the spectra bandwidth, the lower the degree of mineral crystallinity. This disorder changes the selection rules of the Raman active modes: more phonon modes become Raman active, and each phonon mode broadens its features.[40] The slope of the phonon dispersion curves of the vibrational modes determines the shift of the bands:[41] a negative slope results in a shift towards lower wavenumbers.

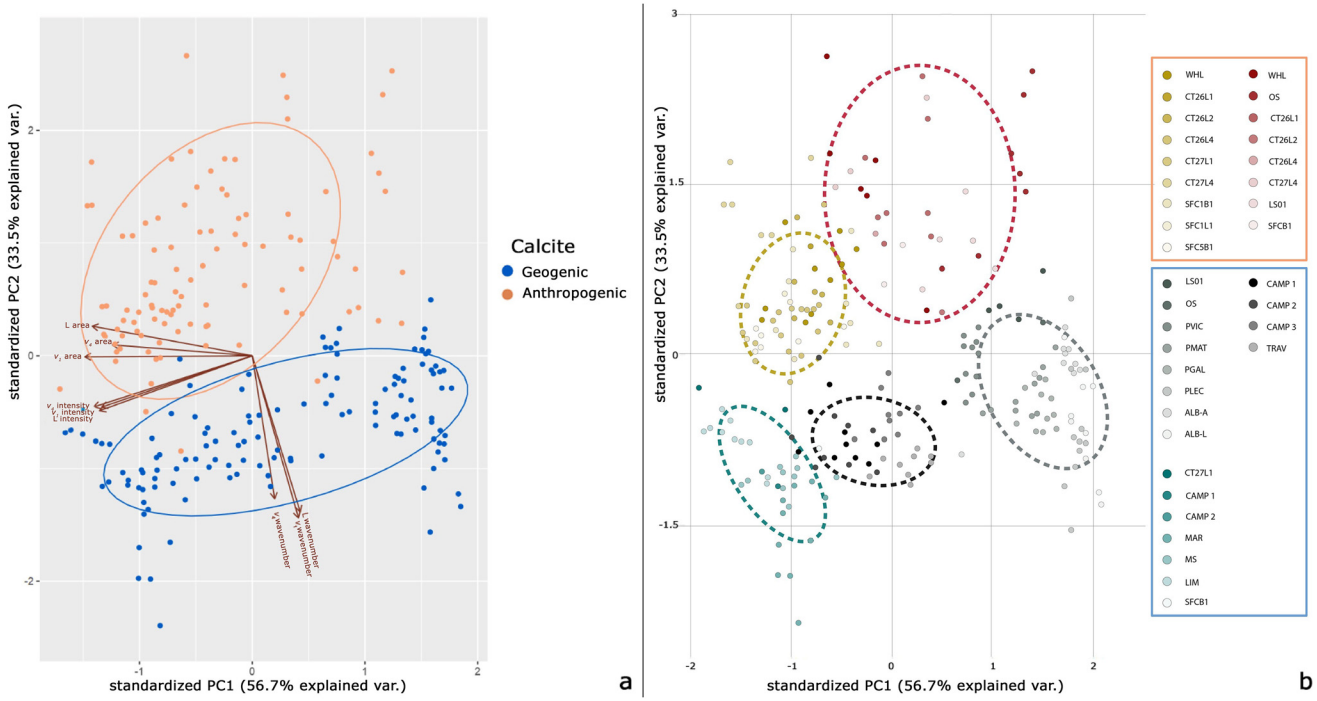

Fig. 6 Biplot from the PCA analysis (a) and K-means clustering plot after performing PCA (b).

Table 3 Classification report of the logistic regression and random forest model performances

Validation methods	Calcite	Precision	Recall	F1-score	Support
Logistic regression	Anthropogenic	0.95	0.97	0.96	38
	Geogenic	0.96	0.93	0.95	29
Accuracy				0.96	67
Macro avg		0.96	0.95	0.95	67
Weighted avg		0.96	0.96	0.96	67
Random forest	Anthropogenic	0.90	0.97	0.94	38
	Geogenic	0.96	0.86	0.91	29
Accuracy				0.93	67
Macro avg		0.93	0.92	0.92	67
Weighted avg		0.93	0.93	0.92	67

As a consequence, calcite disordered systems reflect a systematic larger FWHM and a shift of certain modes toward lower wavenumbers when compared with the crystalline counterparts, as already observed in the literature also for other minerals.[20,42,43]

Crystallinity can vary for many reasons, mainly regarding the increase in crystal defects or in amorphous or nanocrystalline phases in the sample.[44] During the carbonation process, the degree of structural order of calcite can be compromised by several factors, *i.e.* the time of setting and environmental conditions, and there is a general agreement in the literature on the possibility of achieving complete carbonation.[45] The findings achieved in this study highlight that the carbonation process leads to the formation of structurally disordered calcite crystals.

The further outcome of the present study concerns the discriminating power of v_4 intensity within the geogenic samples: sedimentary carbonates, travertine and metamorphic rocks are distinguishable by following their v_4 intensity. The average calcite crystal size is higher in geogenic samples than in anthropogenic samples, and thus polarization effects in Raman spectra of rocks should be considered. The crystallographic orientation of calcite with respect to the incident light polarization is one of the key factors responsible for the relative intensity ratio change of bands in calcite Raman spectra. Interestingly, the outcomes of the present study highlight that mineral orientation depends on the growth and deformation processes of crystals during diagenesis and Raman spectroscopy can be used to distinguish the preferred mineral orientation.[46] The polarization effect is negligible in anthropogenic calcite due to the reduced crystal size, as confirmed by the quite homogeneous v_4 intensity.

Table 4 Full-widths at half-maximum (FWHMs) recorded for L, v_4 and v_1 bands

	L FWHM	v_4 FWHM	v_1 FWHM
Geogenic calcite			
MAR	12.8	5.6	4.7
CAMP 1	11.8	5.3	4.5
CAMP 2	13.4	5.6	4.9
CAMP 3	11.9	5.2	4.5
MS	12.6	5.4	4.6
LIM	15.9	6.8	5.1
PLEC	16.8	5.1	4.3
ALB L	13.2	5.1	4.4
ALB A	17.4	5.8	4.8
TRAV	13.7	5.9	4.5
PGAL	13.1	5.3	4.5
PMAT	13.3	5.6	4.4
PVIC	12.9	5.4	4.5
Anthropogenic calcite			
OS	25.3	8.7	6.7
LS01	22.2	8.8	5.2
WHL	21.9	7.5	6.3
CT26L1	24.3	7.2	5.9
CT26L2	20.2	6.5	5.3
CT26L4	19.3	6.6	5.4
CT27L4	26.6	7.5	6.4
CT27L1	18.1	6.3	5.2
SFC1B1	21.4	7.6	6.0
SFC1L1	20.6	7.6	6.2
SFC5B1	18.5	7.0	5.5

Conclusions

In this work, for the first time, the distinction between geogenic and anthropogenic calcite was made using high-resolution micro-Raman spectroscopy in a non-destructive way. The observed systematic Raman shifts in L and v_1 bands for anthropogenic calcite prompted us to apply data analysis and integrated machine learning methods. The successful parameters (among the position of the band, the intensity of the band, the area subtended by the bands and the FWHMs of L, v_4, and v_1) for distinguishing the calcite origins were identified from KIF, PCA, K-means clustering, and the relationship between the target and the predictor variables using the logistic regression and random forest models.

The proposed method was shown to be effective in discriminating anthropogenic calcite in pyrotechnological materials (*i.e.* mortars and plasters) in order to select the most suitable carbonate fraction for radiocarbon dating purposes. The application of this approach could be extended to the evaluation of the carbonate origins of pigments and sediments in archaeological contexts. Furthermore, types of carbonate rocks are distinguishable on the basis of the v_4 intensity, paving the way to other geological and petrographic applications. The measurements could also be potentially performed *in situ* using portable Raman instruments with a suitable spectral resolution.

The structurally ordered–disordered, crystallinity degree and the polarization effect are the main factors that influence the Raman spectral signature of calcite. These findings encourage further investigations, *i.e.* with single crystal X-ray diffraction to obtain detailed information about the crystal structure of the different examined calcites, extending this research also to the precipitation of secondary calcite in different fields.

Author contributions

Calandra S.: methodology, experiments and data collection, data analysis, writing, review and editing. Conti C.: methodology, data collection and acquisition, and writing and review. Centauro I.: data analysis, system design and development, and writing; Cantisani E.: methodology, project planning, and writing and review.

Conflicts of interest

There are no conflicts to declare.

References

1 G. Artioli, M. Secco and A. Addis, in *The Contribution of Mineralogy to Cultural Heritage*, EMU Notes Miner, 2019, vol. 20, pp. 151–202. DOI: 10.1180/EMU-notes.20.4.
2 E. Cantisani, F. Fratini and E. Pecchioni, *Minerals*, 2022, **12**(1), 41, DOI: 10.3390/min12010041.
3 V. Chu, L. Regev, S. Weiner and E. Boaretto, *J. Archaeol. Sci.*, 2008, **35**(4), 905–911, DOI: 10.1016/j.jas.2007.06.024.
4 L. Regev, K. M. Poduska, L. Addadi, S. Weiner and E. Boaretto, *J. Archaeol. Sci.*, 2010, **37**(12), 3022–3029, DOI: 10.1016/j.jas.2010.06.027.
5 S. Calandra, E. Cantisani, B. Salvadori, S. Barone, L. Liccioli, M. Fedi and C. A. Garzonio, *J. Phys.: Conf. Ser.*, 2022, **2204**(1), 012048, DOI: 10.1088/1742-6596/2204/1/012048.
6 H. G. Machel, R. A. Mason, A. N. Mariano and A. Mucci, in *Luminescence Microscopy and Spectroscopy: Qualitative and Quantitative Applications*, Soc. Sedim. Geol., 1991, vol. 25, pp. 37–57. DOI: 10.2110/scn.91.25.0009.
7 A. El Ali, V. Barbin, G. Calas, B. Cervelle, K. Ramseyer and J. Bouroulec, *Chem. Geol.*, 1993, **104**(1–4), 189–202, DOI: 10.1016/0009-2541(93)90150-H.
8 D. K. Richter, T. Götte, J. Götze and R. D. Neuser, *Mineral. Petrol.*, 2003, **79**, 127–166, DOI: 10.1007/s00710-003-0237-4.
9 M. B. Toffolo, G. Ricci, L. Caneve and I. Kaplan-Ashiri, *Sci. Rep.*, 2019, **9**(1), 16170, DOI: 10.1038/s41598-019-52587-7.
10 M. B. Toffolo, G. Ricci, R. Chapoulie, L. Caneve and I. Kaplan-Ashiri, *Radiocarbon*, 2020, **62**(3), 545–564, DOI: 10.1017/RDC.2020.21.
11 R. S. Krishnan, *Proc. Indiana Acad. Sci.*, 1945, **A22**, 182–193.
12 D. Krishnamurti, *Proc. Indiana Acad. Sci.*, 1957, **A46**, 183–202.
13 W. D. Bischoff, S. K. Sharma and F. T. MacKenzie, *Am. Mineral.*, 1985, **70**(5–6), 581–589.
14 I. P. Herman and F. Magnotta, *J. Appl. Phys.*, 1987, **61**(11), 5118–5128, DOI: 10.1063/1.338286.

15 K. E. Kuebler, A. Wang, K. Abbott and L. A. Haskin, *Lunar and Planetary Science XXXII*, 2001, p. 1889.
16 M. De La Pierre, C. Carteret, L. Maschio, E. André, R. Orlando and R. Dovesi, *Chem. Phys.*, 2014, **140**(16), 164509, DOI: **10.1063/1.4871900**.
17 L. Borromeo, U. Zimmermann, S. Andò, G. Coletti, D. Bersani, D. Basso and E. Garzanti, *J. Raman Spectrosc.*, 2017, **48**(7), 983–992, DOI: **10.1002/jrs.5156**.
18 S. H. Urashima, M. Morita, S. Komatani and H. Yui, *Anal. Chim. Acta*, 2023, 340798, DOI: **10.1016/j.aca.2023.340798**.
19 E. Zolotoyabko, E. N. Caspi, J. S. Fieramosca, R. B. Von Dreele, F. Marin, G. Mor and Y. Politi, *Cryst. Growth Des.*, 2010, **10**(3), 1207–1214, DOI: **10.1021/cg901195t**.
20 U. Wehrmeister, D. E. Jacob, A. L. Soldati, N. Loges, T. Häger and W. Hofmeister, *J. Raman Spectrosc.*, 2011, **42**(5), 926–935, DOI: **10.1002/jrs.2835**.
21 C. Giacovazzo, *Fundamentals of Crystallography*, International Union of Crystallography, Oxford Univ., Press, 3rd edn, 2011.
22 P. Urbanová, E. Boaretto and G. Artioli, *Radiocarbon*, 2020, **62**(3), 503–525, DOI: **10.1017/RDC.2020.43**.
23 A. Lindroos, Å. Ringbom, J. Heinemeier, I. Hajdas and J. Olsen, *Radiocarbon*, 2020, **62**(3), 565–577, DOI: **10.1017/RDC.2020.5**.
24 L. M. Seymour, J. Maragh, P. Sabatini, M. Di Tommaso, J. C. Weaver and A. Masic, *Sci. Adv.*, 2023, **9**(1), eadd1602, DOI: **10.1126/sciadv.add1602**.
25 Z. M. Khoshhesab, in *Infrared spectroscopy-Materials science, engineering and technology*, IntechOpen, 2012, pp. 233–244. DOI: **10.5772/2055**.
26 C. Indelicato, I. Osticioli, J. Agresti, D. Ciofini, A. Mencaglia, M. Perotti and S. Siano, *Eur. Phys. J. Plus*, 2022, **137**(3), 359, DOI: **10.1140/epjp/s13360-022-02536-7**.
27 J. M. Palma-Ruiz, A. Torres-Toukoumidis, S. E. González-Moreno and H. G. Valles-Baca, *Heliyon*, 2022, e08959, DOI: **10.1016/j.heliyon.2022.e08959**.
28 J. Hao and T. K. Ho, *J. Educ. Behav. Stat.*, 2019, **44**(3), 348–361, DOI: **10.3102/1076998619832248**.
29 H. Vinutha, B. Poornima and B. Sagar, in *Information and Decision Sciences. Advances in Intelligent Systems and Computing*, ed. S. Satapathy, J. Tavares, V. Bhateja and J. Mohanty, Springer, Singapore, 2018, p. 701. DOI: **10.1007/978-981-10-7563-6_53**.
30 F. K. Sufi, *Software Impacts*, 2022, **11**, 100218, DOI: **10.1016/j.simpa.2022.100218**.
31 C. Zhu, C. U. Idemudia and W. Feng, *Inform. Med. Unlocked*, 2019, **17**, 100179, DOI: **10.1016/j.imu.2019.100179**.
32 X. Fan, W. Ming, H. Zeng, Z. Zhang and H. Lu, *Analyst*, 2019, **144**(5), 1789–1798, DOI: **10.1039/C8AN02212G**.
33 A. Amjad, R. Ullah, S. Khan, M. Bilal and A. Khan, *Vib. Spectrosc.*, 2018, **99**, 124–129, DOI: **10.1016/j.vibspec.2018.09.003**.
34 Y. Liu, Y. Wang and J. Zhang, in *Information Computing and Applications*. ICICA 2012. Lecture Notes in Computer Science, ed. B. Liu, M. Ma and J. Chang, Springer, Berlin, Heidelberg, 2012, vol. 7473, DOI: **10.1007/978-3-642-34062-8_32**.
35 R. Couronné, P. Probst and A. L. Boulesteix, *BMC Bioinf.*, 2018, **19**, 270, DOI: **10.1186/s12859-018-2264-5**.
36 K. Nakamoto, J. Fujita, S. Tanaka and M. Kobayashi, *J. Am. Chem. Soc.*, 1957, **79**(18), 4904–4908.
37 W. B. White, in *The Infrared Spectra of Minerals*, Mineralogical Society of Great Britain and Ireland, The Carbonate Minerals, 1974, vol. 4, pp. 227–284. DOI: **10.1180/mono-4.12**.
38 B. E. Scheetz and W. B. White, *Am. Mineral.*, 1977, **62**, 36–50.
39 D. Wang, L. M. Hamm, R. J. Bodnar and P. M. Dove, *J. Raman Spectrosc.*, 2012, **43**(4), 543–548, DOI: **10.1002/jrs.3057**.
40 Q. Wang, D. D. Allred and L. V. Knight, *J. Raman Spectrosc.*, 1995, **26**(12), 1039–1043, DOI: **10.1002/jrs.1250261204**.
41 C. L. Jiang, W. Zeng, F. S. Liu, B. Tang and Q. J. Liu, *J. Phys. Chem. Solids*, 2019, **131**, 1–9, DOI: **10.1016/j.jpcs.2019.03.011**.
42 B. Xu and K. M. Poduska, *Phys. Chem. Chem. Phys.*, 2014, **16**(33), 17634–17639, DOI: **10.1039/c4cp01772b**.
43 J. Perrin, D. Vielzeuf, D. Laporte, A. Ricolleau, G. R. Rossman and N. Floquet, *Am. Mineral.*, 2016, **101**(11), 2525–2538, DOI: **10.2138/am-2016-5714**.
44 R. F. Perez and J. Martinez-Frias, *J. Raman Spectrosc.*, 2003, **34**(5), 367–370, DOI: **10.1002/jrs.1003**.
45 J. I. Alvarez, R. Veiga, S. Martínez-Ramírez, M. Secco, P. Faria, P. N. Maravelaki and J. Válek, *Mater. Struct.*, 2021, **54**(2), 63, DOI: **10.1617/s11527-021-01648-3**.
46 A. E. Murphy, R. S. Jakubek, A. Steele, M. D. Fries and M. Glamoclija, *J. Raman Spectrosc.*, 2021, **52**(6), 1155–1166, DOI: **10.1002/jrs.609**.