Research article

Journal of RAMAN SPECTROSCOPY

Received: 1 October 2014 Revised: 17 June 2015 Accepted: 26 June 2015 Published online in Wiley Online Library: 10 August 2015

(wileyonlinelibrary.com) DOI 10.1002/jrs.4757

Machine learning tools for mineral recognition and classification from Raman spectroscopy

C. Carey,[a]* T. Boucher,[a] S. Mahadevan,[a] P. Bartholomew[b] and M. D. Dyar[c]

Tools for mineral identification based on Raman spectroscopy fall into two groups: those that are largely based on fits to diagnostic peaks associated with specific phases, and those that use the entire spectral range for multivariate analyses. In this project, we apply machine learning techniques to improve mineral identification using the latter group. We test the effects of common spectrum preprocessing steps, such as intensity normalization, smoothing, and squashing, and found that the last is superior. Next, we demonstrate that full-spectrum matching algorithms exhibit excellent performance in classification tasks, without requiring time-intensive dimensionality reduction or model training. This class of algorithms supports both vector and trajectory input formats, exploiting all available spectral information. By combining these insights, we find that optimal mineral spectrum matching performance can be achieved using careful preprocessing and a weighted-neighbors classifier based on a vector similarity metric.

Keywords: Raman; mineral identification; spectral library search; machine learning

Introduction

Use of Raman spectroscopy in the geosciences is growing rapidly, as evidenced by burgeoning publications in the fields of geology, cultural anthropology, environmental science, and, in particular, planetary science. Raman spectrometers have been proposed for exploration of a diverse range of extraterrestrial targets including asteroids,[1] Europa,[2,3] Mars,[4–6] the Moon,[7] and Venus.[8] A Raman laser spectrometer (RLS) is part of the science instrument payload of the European Space Agency 2018 ExoMars mission; the RLS instrument will target mineralogical and astrobiological investigations on the surface and subsurface of Mars.[9,10] Raman will also be used on the upcoming NASA Mars 2020 mission as part of the SuperCam and SHERLOC instruments.[11] What all these applications have in common is their dependence on software and mineralogical databases for phase identification and quantification of relative abundances of mineral components. Because of the structural diversity and chemical complexity of naturally occurring minerals, optimal applications for these purposes require an infusion of work into development of appropriate software and mineral databases.

In practice, users commonly depend on a combination of matching software distributed by spectrometer manufacturers and one-on-one comparisons with database spectra for their identifications, but these types of identification have two critical limitations. First, identifications are only as good as the databases used to match them. No database can be entirely comprehensive, so it cannot be unequivocally claimed that any spectrum is a 'perfect match' to exactly one other phase.

The RRUFF database was founded in 2006 at Arizona State University by Robert Downs to remedy this situation by providing coverage of all known mineral species.[12] It has quickly become the preeminent resource available for Raman spectra of minerals. RRUFF currently contains over 20 378 spectra acquired at several different laser wavelengths from oriented and un-oriented samples, and a new user interface is currently under design. However, its weakness is that many of its mineral species identifications have not yet been confirmed using independent techniques such as X-ray diffraction (XRD), and its spectra are of highly variable quality, many with mild fluorescence that can interfere with Raman spectral matches. Although a ratings system exists to guide users to the highest quality spectra on RRUFF, it is rarely used, and the RRUFF project's CrystalSleuth[13] matching software defaults to querying a database in which only 64% of samples have XRD confirmation. Moreover, RRUFF by design focuses on representing a broad range of the >4000 known mineral species rather than on presenting multiple examples of common rock-forming species, so spectral matches with exceedingly rare minerals can occur if users are not geologically astute.

The second current limitation to mineral identification is that users commonly depend on matching software, both CrystalSleuth and other proprietary products, for identifications. These tools fall into two groups: those that are largely based on fits to diagnostic peaks associated with specific minerals and those that use the entire spectral range for multivariate analyses. In geological studies, both types of tools are challenged by mixtures of minerals in fine-grained or powdered rock samples, and by the availability and quality of Raman databases for minerals. Moreover, existing search/match software packages tend to have several issues. In spectra with one strong peak and several distinct but low-intensity peaks, high match weighting is given to the strongest peak(s), and smaller peaks may be virtually ignored.

* *Correspondence to: C. Carey, Department of Computer Science, University of Massachusetts at Amherst, Amherst, MA, USA.*
E-mail: ccarey@cs.umass.edu

a *Department of Computer Science, University of Massachusetts - Amherst, Amherst, MA, USA*

b *Department of Astronomy, University of New Haven, West Haven, CT, USA*

c *Department of Biology and Environmental Science, Mount Holyoke College, South Hadley, MA, USA*

Journal of RAMAN SPECTROSCOPY

Noisy spectra with low peak intensities may produce high match scores to other noisy spectra in the reference database because they happen to have a similar pattern to variations driven by random noise. Existing search/match software is computationally cumbersome, resulting in long run times when the reference database is large. Finally, spectra present several different styles and shapes of peaks, as well as degrees of photoluminescent interference.

Given this situation, the accuracy of existing tools for mineral identification is in need of improvement. For example, the CrystalSleuth fingerprinting software used predominantly in this community is only ~84% accurate (Robert Downs, personal communication, 2014) for single mineral species identification using only the highest quality RRUFF data, and it does not detect minor phases included in mineral mixtures. Our current research effort seeks to improve this overall situation by improving the software available for mineral identification in single minerals and adding additional geologically relevant samples including mineral mixtures to the new NASA Geosciences Planetary Data System implementation of RRUFF (set to debut in 2015). This paper focuses on the former issue by using a carefully selected subset of 3950 samples of RRUFF data to evaluate full-spectrum matching algorithms using combinations of preprocessing steps, spectrum similarity measures, and neighbor-based classifiers, which show great promise in applications involving mineral identification.

The following sections cover related work on spectrum matching problems then describe the data and methods used in this study. These methods consist of choices about preprocessing, classification, and spectrum similarity metrics, each of which are examined independently to characterize their effect on overall performance. Experimental results are presented and discussed, including a comparison with recent work on this problem. We conclude with recommendations for optimal matching based on these results.

Background

Several workers have considered the issues involved with automatic matching and identification of Raman and other types of spectra, with broad-ranging application domains. While early efforts relied on expert knowledge of spectral features, more recent approaches have made use of a wide range of statistical and machine learning tools. For example, discriminant analysis was used to identify additives in honey using Fourier Transform-Raman spectroscopy.[14] Support vector machines have been used broadly, with applications including mineral detection with near-infrared spectroscopy[15] and composition prediction with Raman spectra.[16] Artificial neural networks (ANNs) have seen increasing use in spectroscopic applications as well.[17–19] Similarity-based methods for both peak-feature[6] and full-spectrum matching[20,21] have also been explored, especially for Raman spectroscopy. Some approaches restrict the task to identification of a specific component,[14,15] while others attempt to cluster spectra into logical groups.[19,22] Most methods employ some combination of spectrum preprocessing steps to reduce the influence of noise and fluorescence,[23,24] and some also project spectra into a lower-dimensional feature space, typically using principal components analysis (PCA).[22,25] These applications focus on single-phase identification, although some workers have proposed algorithms for determining the relative contributions of phase in simple mixtures.[26–28]

Specifically in the geosciences, pioneering work in automated mineral identification has been performed by Pablo Sobron and colleagues involved with the RLS instrument on ExoMars,[6,18,26] building upon earlier technique development at Washington University by Larry Haskin and Alian Wang.[5,29] These studies focus on automated identification of minerals using univariate analysis,[6,30–32] but they are not fully adaptable to mineral mixtures. A recent paper by Lopez-Reyes *et al.*[18] explores the use of PCA, partial least squares, and ANNs to quantify sulfate mineral abundances in binary mixtures from laser Raman spectral data. In their experiments on a small number of samples (17), they demonstrated that ANN models of mineral mixtures not only provide 100% detection accuracy but also predict abundances of the components.

Although these previous studies laid the groundwork for mineral identification in pure phases, none has utilized the full RRUFF Raman data set as a training set and then tested algorithms against that large number of spectra. The accuracy of mineral species and group identifications using RRUFF data has not been fully assessed, nor have modern autonomous learning techniques developed for analogous applications been applied to this domain. The lack of such previous work forms the underlying motivation for the current study, which approaches this task from the perspective of the machine learning community, in which many state-of-the-art techniques suitable for this problem have been developed.

Data used

For this project, we employed a subset of spectra from the RRUFF database, using only data from un-oriented samples collected at random orientations. We downloaded spectra of these samples in preprocessed form from RRUFF, with baseline correction and instrumental artifacts removal already applied. Spectra with overwhelming specimen fluorescence were excluded, as were spectra without an independent mineral species identification from XRD analysis. Each mineral name was matched with its four-part Dana classification number.[33] Samples without Dana numbers were excluded. The final set of spectra contained 3950 spectra, representing 1215 different mineral species over 78 Dana classes.

Each spectrum used in our study was baseline corrected by the RRUFF project's algorithm, which uses piecewise linear interpolation between smoothed off-peak segments.[34] However, we note here that several other techniques have been proposed to automatically subtract baselines from spectra,[35–39] each with tunable parameters that may be set to refine the quality of the removed baseline. In practice, it is common to find empirically a set of parameters that produce reasonable results on a small set of spectra then use that setting to process the remaining samples in the spectral library. Because no baseline correction algorithm is perfect and users cannot find perfect parameter settings for all data, the continuum removal process typically introduces systematic error to the inputs of a spectrum matching algorithm. It follows that good matching algorithms should not be sensitive to the types of error that baseline correction creates.

In this paper, we use RRUFF's baseline-subtracted data, for which the associated errors are as yet untested. We suspect that future work may show improvements in spectral identification accuracy when baseline correction is integrated within the matching process, and this is an area of active pursuit in our research group. However, because a majority of the Raman community in the geosciences uses RRUFF data and CrystalSleuth software without modification, the present study intentionally employs only standard RRUFF processed data.

Table 1. Summary of results

	Average accuracy (%)	Class	Type	Group	Species
Vector sim.	1-NN, norm[a]	**94.9**	92.9	90.7	82.1
	10-NN, norm	93.8	91.7	89.9	82.2
	WN, norm	93.1	91.2	89.7	82.3
	WN, sqrt+norm	93.4	91.9	90.7	83.5
	WN, sqrt+norm+sigmoid	94.8	**93.3**	**92.0**	**84.8**
Trajectory sim.	1-NN, norm	83.3	80.9	78.3	69.1
	10-NN, norm	84.6	82.2	79.6	70.8
	WN, norm	84.7	82.5	79.9	71.0
	WN, sqrt+norm	90.3	88.0	85.6	77.2
	WN, sqrt+norm+sigmoid	86.5	84.3	81.7	73.3
	WN, sqrt+norm+sqrt	**92.1**	**89.1**	**86.8**	**78.0**
	Multilayer perceptron	59.0	51.9	46.0	35.6
	Decision tree	44.3	40.5	37.3	31.6

k-NN and WN refer to nearest neighbors and weighted-neighbors classifiers, respectively. Preprocessing steps are abbreviated as follows: 'norm' for maximum intensity normalization, 'sqrt' for square root squashing, and 'sigmoid' for sigmoid squashing. The best results for vector and trajectory similarity matching are highlighted in bold.
[a]Equivalent to CrystalSleuth's matching method.

Spectral matching techniques

In geoscience applications, spectral matching is particularly complicated by variations in chemical composition of different minerals. For example, the mineral species albite and anorthite are end-members of a chemical solid solution between sodium and calcium feldspar, the most commonly occurring mineral group in the Earth's crust. However, naturally occurring feldspars cover the entire compositional range between 100% albite and 100% anorthite, and each increment may have its own Raman signature. For example, this is observed in olvine group minerals.[40]

Moreover, accuracy in spectral matching should be qualified in terms of the geoscience user's goals. For example, a user might place the value of classifying a sample into a mineral group nearly equal to classifying to the species level. The Dana classification system is the most widespread hierarchical descriptor of minerals. Each Dana number has four parts, denoting mineral class, type, group, and species. As an example, the species albite (76.01.03.01) is a member of the plagioclase series in the feldspar group (76.01.03), in the Al-Si framework type (76.01), and the tectosilicate Al-Si class (76). A positive match with anorthite (76.01.03.06) would be viewed as incorrect at the species level, but accurate at the group, type, and class levels. In many cases, geoscience users of spectral matching software may not fully appreciate the complexity of this nomenclature, which can lead to confusing and inconsistent reporting of matching results. Success in matching minerals from Raman spectroscopy must therefore be carefully defined to be at a specific level, e.g. species and group, as is performed in this project (Table 1).

Preprocessing

For many spectrum matching methods, all spectra in the data set need to be sampled at a common set of wavelengths. The RRUFF set was collected from heterogenous sources, and as a result, individual spectra have different sampling rates and measurement ranges. Thus, in cases where resampling was necessary, simple linear interpolation was used to convert each spectrum to a vector of 1715 intensity values, sampling uniformly from 85 to 1800 cm^{-1}. This range of wavenumbers was chosen to maximize the number of RRUFF spectra that could be used in our analyses. For samples that did not cover the entire range of wavelengths, zeros were supplied in place of missing intensity values. Linear interpolation was employed because the majority of spectra in the RRUFF set are sampled more frequently than our target rate of once per cm^{-1}, which mitigates the need for clever interpolation techniques.

Because of differences in sample crystal orientation, laser polarization, focus, and other instrumental parameters, peaks in matching spectra of the same species often vary in intensity (Fig. 1), inducing dissimilarity in most matching algorithms. To counteract this issue, we apply several steps of nonlinear, monotonic preprocessing before performing similarity computation, as described in the succeeding text. These preprocessing steps include square root squashing, sigmoid squashing, and various types of intensity normalization. 'Squashing' refers to a transformation function f, which is applied to each wavelength of a spectrum independently to produce a new spectrum with smaller distances between strong and weak spectral features. Square root squashing uses $f(x) = \sqrt{x}$, while sigmoid squashing uses $f(x) = \frac{1-\cos(\pi x)}{2}$. Intensity normalization is the process of scaling each wavelength of a spectrum based on a statistic of the entire spectrum; choices include the maximum value (L_∞ norm), the sum of absolute values (L_1 norm), and the sum of squared values (L_2 norm). Each squashing or normalizing preprocessing step preserves the relative

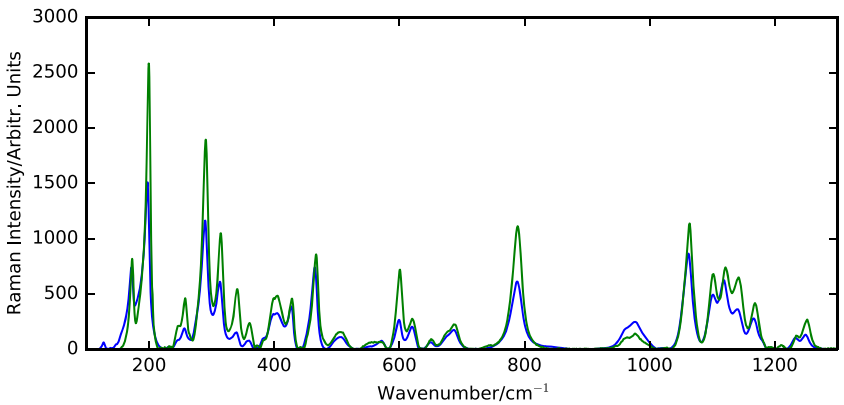


Figure 1. Raman spectra acquired using a 532-nm laser of trolleite samples 30565 (blue) and 32267 (green) from the RRUFF database. Variations in peak intensities occur even in samples of the same mineral species and laser energy.

Journal of RAMAN SPECTROSCOPY

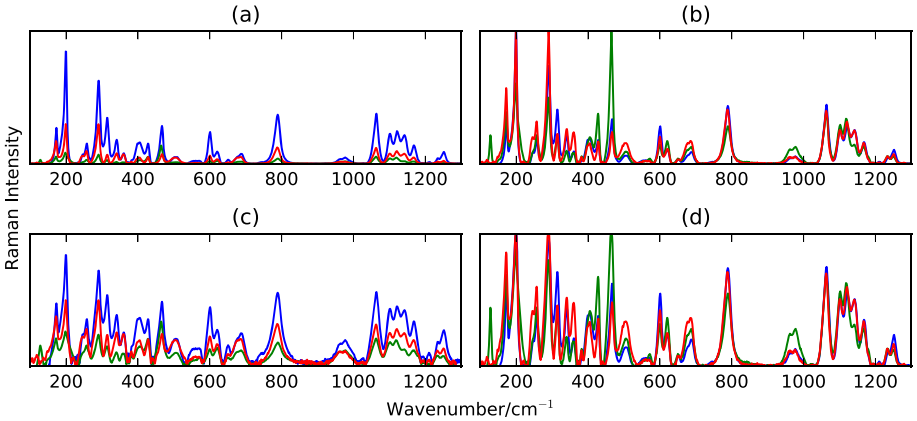


Figure 2. Visualization of the effects of various preprocessing steps. (a) Three resampled, unprocessed spectra of the trolleite mineral species (RRUFF samples 32267 in blue, 30567 in red, and 32265 in green). (b) The same data, rescaled by normalizing to maximum value. (c) Data rescaled using the square root of each wavelength's intensity. (d) A combination of (b) and (c), performing square root squashing, then maximum value normalization, then sigmoid squashing. The last step scales intensity using the sigmoid function $f(x) = \frac{1-\cos(\pi x)}{2}$.

ordering of intensity values while mitigating the effect of peak intensity differences (Fig. 2).

Because spectral data are inherently high-dimensional, methods for reducing dimensionality can be effectively employed. In this study, we first examine the use of PCA, which solves for a linear mapping from a resampled spectra's 1715-dimensional input space to a lower-dimensional feature space while preserving as much variance as possible.[41] However, this approach ignores the sequential nature of spectral data, treating each channel individually. This results in an unfortunate loss of information that could otherwise be exploited by matching algorithms.

We also investigated other preprocessing approaches, such as smoothing and first derivative transforms, using various parameterizations of the Savitzky–Golay filter.[42] Such approaches preserve the sequential nature and shape of the spectral data. As with the squashing and normalizing steps, these preprocessing procedures also serve to eliminate intra-species variation while maximizing intra-species distances, albeit in a more destructive manner due to their lack of monotonicity. Our preliminary experiments showed that these types of preprocessing do not significantly improve upon the steps shown in Fig. 2 for our data, so we chose not to include them in further tests.

Weighted-neighbors classification

The simplest classification scheme is a nearest neighbor classifier, which ranks reference spectra by their similarity with the query spectrum. The identification of the most similar reference spectrum (e.g. Dana species, group, type, and class) is then assigned to the query. This approach is simple and reliable and is used by CrystalSleuth for its library search function. This kind of classification requires a similarity metric that captures the underlying mineralogical relationships between pairs of spectra. This family of classifiers is agnostic to the choice of metric, however, so we defer discussion of that problem to the next section.

In our experiments, we use a variant of the nearest neighbors approach called a 'weighted-neighbors' classifier. In this variant, we rank each mineral species by its average similarity score with all reference spectra identified as the same species. This allows the classifier to use spectrum-level similarity as a proxy for similarity with an entire mineral species, without having to construct a prototypical spectrum for each species, such as a mean intensity spectrum.

Similarity metrics

As noted in the previous text, the success of a full-spectrum classification tool rests primarily on the choice of similarity metric, a function that compares two spectra and produces a real-valued score that corresponds to how similar the inputs are. An ideal metric should capture the mineralogical properties of a Raman spectrum while ignoring the influence of noise, continuum effects, and other perturbations. Most importantly for the identification task, this metric should be able to distinguish between minerals of different species, regardless of sample origin. In this section, we discuss vector- and trajectory-based similarity metrics.

Vector similarity metrics

Vector-based similarity metrics are well studied and widely used, both in spectrum matching contexts and throughout machine learning.[43] Euclidean (L_2) distance is a common example of such a metric. When applied to spectroscopic data, this class of metrics requires that spectral intensity vectors cover a common set of wavelengths across all samples. Thus, cropping and resampling are often necessary, especially when working with data collected from heterogenous sources. In each of our vector similarity experiments, all spectra were cropped and resampled as described in the previous text, resulting in a 1715-dimensional vector representation.

A popular vector similarity metric for full-spectrum matching is cosine similarity, sometimes known as spectral angle similarity. This metric is simple to calculate and tends to work well in practice and forms the core of the library search functionality in the popular CrystalSleuth software.[12] The cosine similarity metric is defined as

$$similarity\,(x, y) = \frac{x \cdot y}{\|x\|\,\|y\|}$$

for two vectors of spectral intensity values x and y. Intensity values from each spectrum are non-negative, so this similarity score will range between zero and one. High similarity values are produced when a pair of spectra has many matching wavelengths with high intensity. If either of the two spectra has low intensity at

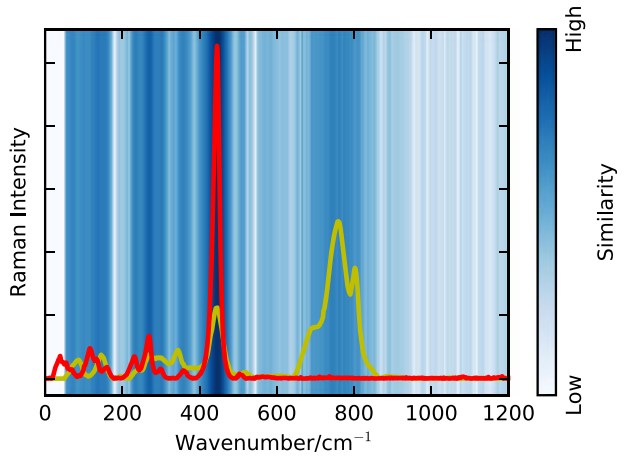


Figure 3. Illustration of cosine similarity. Two resampled 532-nm spectra from chiolite (red) and cabalzarite (yellow) are overlaid on a visualization of each wavelength's individual contribution to the overall cosine similarity score (blue, log-scale). The overlapping pair of peaks in dark blue near 450 cm^{-1} accounts for roughly 85% of the total similarity score.

a particular wavelength, then that pair will not contribute significantly to the overall similarity score at that wavelength. This effect is demonstrated in Fig. 3.

Trajectory similarity

One disadvantage of vector-based approaches is their resampling requirement: all spectral vectors must be sampled in the same interval, at the same wavelengths. To produce strong matches, peaks must fall in exactly the same locations in energy (i.e. the *x*-axis) in every spectrum of a particular species. This requirement is not always met in practice. Compositional substitutions in minerals (such as the solid solution between albite and anorthite mentioned in the previous text) are known to cause peak shifts. In almost any spectrum, there will be peak centroids that lie exactly on the boundary between channels and thus fall sometimes in a higher energy bin, and sometimes in a lower energy bin. Differences in instruments may also cause shifts in peak maxima. Because vector similarity approaches rely on exact correspondence between specific channels, all of these issues can cause problems. For all these reasons, real-world data sets are often heterogeneous, and thus, vector-based methods typically require destructive resampling. In contrast, methods that compute similarity among trajectories are able to operate on spectra directly.

We define a trajectory as an ordered sequence of n wavelength–intensity pairs:

$$T = \{(w_1, intensity_1), ..., (w_n, intensity_n)\}$$

Several algorithms exist for computing similarity between arbitrary trajectories, such as dynamic time warping,[44] longest common sub-sequence (LCSS),[45] and the discrete Frechet distance.[46] These each produce a measure of similarity between two trajectories of arbitrary length, with minimal assumptions about properties of the input data.

To more accurately model similarity for the case of spectrum trajectories, however, adaptations of the generic trajectory algorithms are required. For example, in the spectrum alignment task, extensions to dynamic time warping have been proposed that apply constraints to the produced warping path based on properties of spectroscopic data.[47–49]

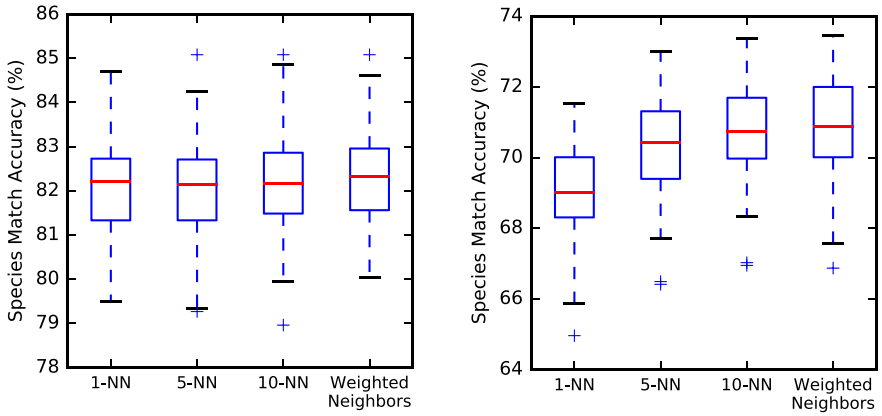
In a similar fashion, we introduce a novel adaptation of LCSS similarity that is specifically designed for spectroscopic data. We employ the minimum bounding envelope approach from the classic LCSS algorithm to compute matching points in a pair of trajectories then add a specialized scoring function that captures pointwise spectral similarity. This algorithm is outlined in Algorithm 1.

Algorithm 1 LCSS-based Spectrum Trajectory Similarity

Input: T^A, T^B spectral trajectories
Parameters: a MatchScore function
$\varepsilon_w = \frac{1}{2}\sqrt{s(T^A)\,s(T^B)}$, where $s(T) = \frac{1}{|T|}\sum_i \left|T_{w_i} - T_{w_{i+1}}\right|$
score $\leftarrow 0$
$k \leftarrow 1$
for $i \in \left[1..\left|T^A\right|\right], j \in \left[1..\left|T^B\right|\right]$ **do**
  $(\delta t_w, \delta t_{int}) \leftarrow \left|T_i^A - T_j^B\right|$
  if $\delta t_w \leq \varepsilon_w$ **then**
    score $\leftarrow$ score + MatchScore $\left(T_i^A, T_j^B\right)$
    $k \leftarrow k + 1$
  end if
end for
Output: similarity $\leftarrow \frac{\text{score}}{k}$

This algorithm calculates a reasonable value for ϵ_w, which controls the 'wavelength looseness' of the match. Intensity comparisons are only made for pairs with absolute wavelength differences within this radius. This aspect of the algorithm allows the computation of similarity between trajectories of different lengths, sampling intervals, and start/end points. If both trajectories are sampled on the same wavelengths, this value could be set to zero such that only exactly matching wavelengths will be compared. In practice, spectra are sampled at slightly

Journal of RAMAN SPECTROSCOPY



(a) Vector similarity.

(b) Trajectory similarity.

Figure 4. Box and whisker plot of weighted-neighbors classifier accuracy, compared with that of nearest neighbor classifiers using one, five, and ten neighbors, respectively. Each classifier used normalization preprocessing, and results were aggregated over 200 randomized trials. In these plots, red lines represent the median accuracy, the box tops and bottoms represent the 75th and 25th percentile accuracies, respectively, and the extent of the dashed lines represents the range of accuracies.

different frequencies, so ϵ_w is computed using each trajectory's mean sampling interval.

The only user-provided parameter is a generic MatchScore function, which compares two wavelength–intensity pairs to produce a scalar similarity value. The choice of this function is critical for accurate spectrum matching, and several options are available. The first is the trajectory analog to cosine matching:

$$\text{MatchScore}\,(x,y) = x_{int} \times y_{int} \tag{1}$$

This will produce the same similarity score as the vector cosine similarity, assuming both trajectories have matching wavelengths and normalized intensities. For our application to spectrum matching, we introduce a slightly more complex scoring function:

$$\text{MatchScore}\,(x,y) = \frac{\min\,(x_{int}, y_{int}) \times}{\left(1 - |x_{int} - y_{int}|^{\alpha}\right)} \tag{2}$$

This function more accurately captures the features of matching spectra: the first term penalizes matches with low intensity, while the second promotes matches with similar intensity values. The amount of promotion for similar values is controlled by a parameter $0 < \alpha \leq 1$.

Experimental results and discussion

We evaluate accuracy of classification using cross validation with semi-randomized splits of the RRUFF data described in the previous text divided into query and reference sets. To ensure that each query spectrum has a true match in the reference set, the reference set for each trial is constructed by selecting three spectra per mineral species at random. The remaining spectra are assigned to the trial's query set. This means that species with three or fewer total spectra are not present in the query set for any trial and always appear in the reference set.

Under this testing regime, each trial selected 2643 reference spectra and 1307 query spectra. Each reference set covered all 1215 mineral species, while each query set contained only the 349 mineral species with more than three samples, representing 62 of the 78 total Dana classes.

Accuracy results were computed by taking the ratio of the number of correct matches to the total number of spectra in the query set, for each level of matching in the Dana hierarchy. Thus, accuracy at the species level can never be greater than accuracy at the group level, and so on.

The CrystalSleuth software was designed only for interactive use, so we could not test against it directly. However, inspection of the CrystalSleuth source code confirmed that it uses a vector similarity classifier. This is equivalent to our experimental setup with a 1-nearest neighbor classifier using vector similarity after intensity normalization preprocessing (row 1 of Table 1).

Choice of classifier

The weighted-neighbors classifier showed marginal improvement over nearest neighbors classifiers (Fig. 4). This effect is likely due to the ability of the weighted-neighbors classifier to overcome

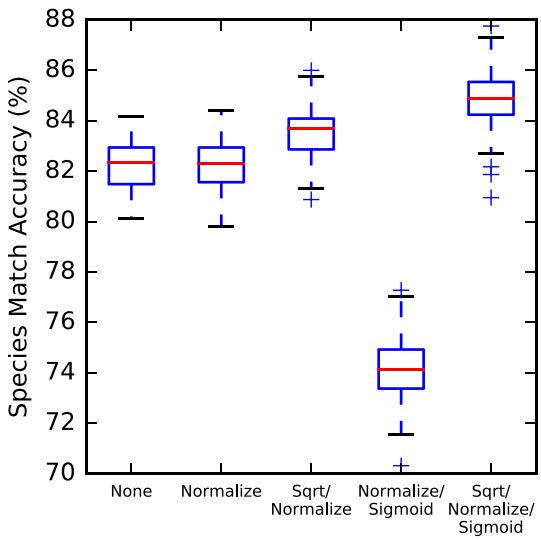


Figure 5. Box and whisker plot of spectrum matching accuracy, with varied preprocessing steps. Colors as described in Fig. 4. Each test used a weighted-neighbor cosine similarity classifier, over 100 randomized trials.

the influence of a single mismatched spectrum in the presence of several true matches with lower similarity scores.

Effect of preprocessing

Figure 5 shows the effect of various preprocessing steps using the vector-based cosine similarity metric and a weighted-neighbors classifier. These tests confirm that the 'squashing' preprocessing steps provide the most effective transformations, although only when used in the correct context and ordering. When sigmoid squashing is used alone, accuracy degrades significantly. When sigmoid squashing is used after square root squashing, it reduces prediction variance without losing accuracy. These results match our expectations based on Fig. 2. We show results only for maximum intensity normalization, because the other proposed normalizers did not offer significant improvement.

Although PCA preprocessing was shown to improve accuracy in a previous study,[22] our results suggest the opposite (Fig. 6). We observed that more principal components resulted in more accurate classification, but that eschewing PCA preprocessing altogether achieved the best performance. This may be due to the size of the spectral library used: our set is roughly ten times larger and has many more mineral species. We also found that the benefits of using PCA were very dependent on other preprocessing steps, especially spectral intensity normalization.

Vector *versus* trajectory similarity

In contrast to vector-based cosine similarity, the new LCSS-based trajectory matching technique is evaluated in Fig. 7. Intensity values were scaled using the same preprocessing steps from the vector similarity experiment in the previous section, but no resampling or cropping was required. This approach fell short of the accuracy achieved by the vector similarity metric, however, so further study will be required to produce an optimal trajectory-based similarity metric. See Table 1 for a comparison of accuracies. The preprocessing steps tested produced changes in relative performance similar to the vector similarity tests, which indicates that the utility of these preprocessing methods is not tied to a specific similarity metric.

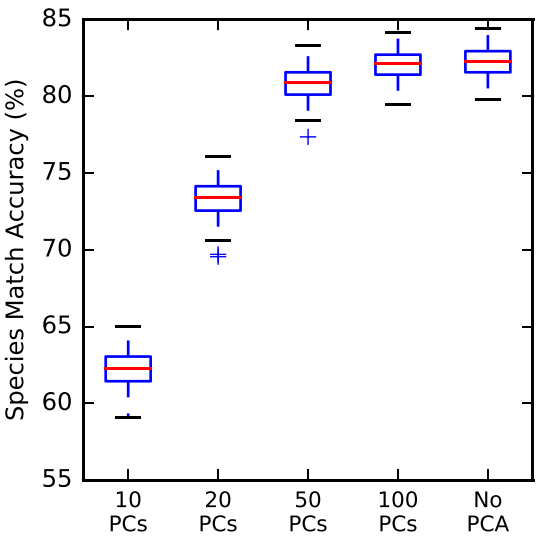


Figure 6. Effects of principal components analysis (PCA) preprocessing. Colors as described in Fig. 4. Each test used a weighted-neighbor cosine similarity classifier, over 100 randomized trials. The *x*-axis shows measured accuracy with varied numbers of principal components (PCs), each with normalized spectral intensity. Accuracy increased as more PCs were used, but direct comparison without any PCA preprocessing performed best. Fully 90% of the spectral library's variance was explained using 50 PCs, which may explain the large drop-off in accuracy below that number. To explain 95% and 99% of variance, 66 and 127 PCs were required, respectively.

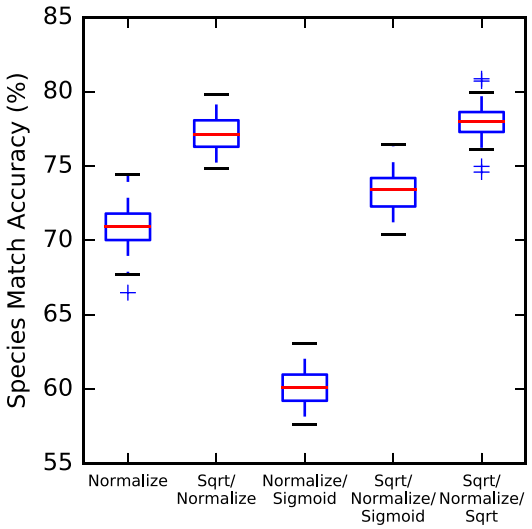


Figure 7. Trajectory similarity results. Colors as described in Fig. 4, with preprocessing methods as described in Fig. 5. For each of 50 randomized trials, the MatchScore function from Eqn (2) was used with parameter $\alpha = 0.1$, selected because it provided the best overall results.

Comparison with a smaller reference data set

Results of this study can be compared against those of Ishikawa and Gulick,[22] who used a much smaller subset of minerals from the RRUFF data set. Their paper lists the mineral species included in their data set but does not specify that individual spectra were used. To create a fair comparison for evaluation, we simply selected all of the species spectra used by Ishikawa from the full RRUFF set, for a total of 214 spectra. We also note that the Ishikawa paper miscategorizes annite in the plagioclase group, but we made no attempt to correct for the influence of that error.

The best algorithm presented by Ishikawa obtains an overall prediction accuracy of 80.4% on their RRUFF subset when matching at the group level. For comparison, we implement both the decision tree and multilayer perceptron classifiers as described in their 2013 paper. For both classifiers, we apply max-normalization and square root squashing then used the first 59 principal components, explaining 99% of the variance (Fig. 8).

Using the cosine similarity metric with a weighted-neighbors classifier, we achieve an overall prediction accuracy of 97.8% at the group level and 89.2% at the species level. These accuracies demonstrate that simple full-spectrum matching techniques can achieve equivalent or better accuracy than sophisticated neural-network classifiers, with a fraction of the computing resources. For example, our technique does not require PCA preprocessing and avoids costly classifier training and tuning steps. When applied to the full RRUFF data set, the multilayer perceptron classifier took 7.7 times longer than our technique, and the decision tree classifier took 18.6 times longer. In addition to speed,

Journal of RAMAN SPECTROSCOPY

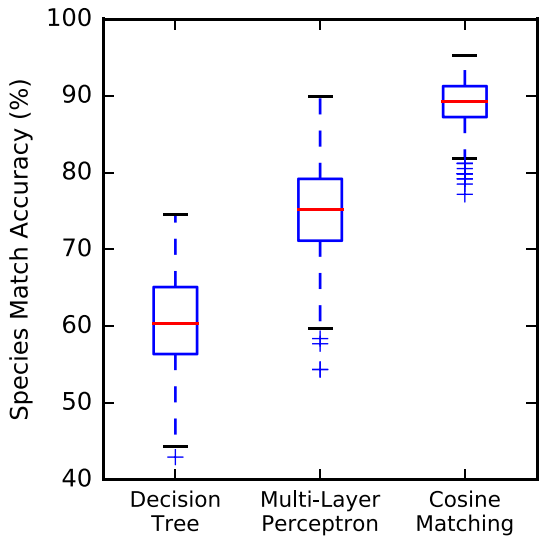


Figure 8. Classifier comparison on the Ishikawa subset, with 500 randomized trials per classifier. Colors as described in Fig. 4. The decision tree classifier achieved 75.0% group-level accuracy on average, with 60.5% accuracy at the species level. The multilayer perceptron classifier averaged 90.3% group-level accuracy, but only 74.9% species-level accuracy. These results mirror the results reported by Ishikawa, despite the slight differences in the set of spectra used. In contrast, our weighted-neighbors cosine similarity classifier achieved 97.8% group-level accuracy and 89.2% species-level accuracy on average, without any model training or parameter tuning.

this reduction of tunable parameters also simplifies the classification process from an end-user standpoint. The results from testing these methods on the full data set are presented in Table 1.

While our experiments with PCA preprocessing step show limited utility on the large RRUFF library (Fig. 6), we find that PCA is more effective on the smaller Ishikawa mineral subset. This implies that the minerals selected by Ishikawa are more linearly separable, and thus more easily classified. This notion is supported by the observed increase in species-level accuracy of the weighted-neighbors classifier when applied to the smaller spectra library. This separability is also demonstrated in Fig. 9, which provides a visual representation of similarity across the Dana hierarchy.

Conclusions: recommendations for optimal matching

This study demonstrates that full-spectrum matching algorithms exhibit excellent performance in classification tasks without expensive dimensionality reduction or model training. This class of algorithms supports both vector and trajectory input formats, exploiting all available spectral information. We introduce a novel trajectory similarity measure designed specifically for heterogeneous spectrum matching and show that it trades some mineral identification accuracy for increased flexibility. We believe that the observed shortcomings in species identification accuracy using trajectory methods will be overcome as the methods are explored further and the details of matching with heterogeneous sampling rates and intervals are optimized.

These techniques offer great promise for adaptation to planetary and terrestrial applications, and our results demonstrate the potential for expanding the application of full-spectrum algorithms to a wide variety of larger data spectral sets and applications for other kinds of materials, as well as mineral mixtures.

Based on the results of this study as summarized in Table 1, we recommend that optimal mineral spectrum matching performance can be achieved using a weighted-neighbors classifier based on a vector similarity metric, on spectra that have been individually preprocessed via square root squashing, then maximum intensity normalization, then sigmoid squashing.

We suspect that even better performance is achievable when using larger, more consistent spectral libraries. We recommend that library development prioritize acquisition of clean spectra without fluorescence (such as are obtained using time-resolved Raman), and consistent measurement parameters (spectral range and resolution). As the community expands its efforts to grow Raman databases, it would be beneficial to come to a consensus on these parameters. Furthermore, it is essential that every mineral species represented in a Raman reference database be correctly identified with X-ray diffraction and chemically characterized. When inconsistencies among spectra of the same species are observed, spectral matching cannot work reliably.

The goal of this project is to improve the software available for mineral identification in single minerals, in order to lay the

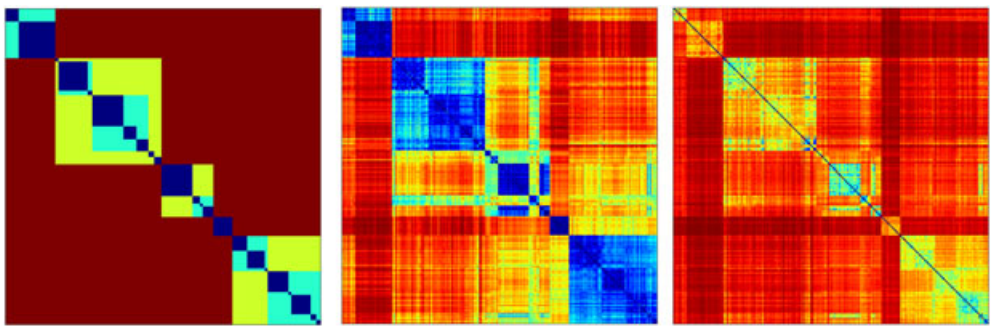


Figure 9. Intraclass similarity of the Ishikawa mineral subset. Each plot represents pairwise similarity between all 216 samples, with colors ranging from blue (high similarity) to red (low similarity). The similarity between any spectrum *i* and spectrum *j* is displayed in row *i* and column *j*. Spectra have been ordered such that samples of the same Dana group and species are adjacent. (Left) Ground truth, in which yellow denotes a class-level match, cyan denotes a group-level match, and blue denotes species-level match. (Middle) Preprocessed cosine similarity. (Right) Preprocessed trajectory similarity. Comparing the ground-truth similarities with the experimental cosine similarity metric shows that many mineral species are distinctly separated, and that even misclassified spectra are likely to be within the correct Dana class. A similar result is visible in the trajectory similarity matrix, although individual Dana species are harder to distinguish.

groundwork for the study of mineral mixtures. We find that the available spectral libraries limit the accuracy of matching methods, but improvement in existing matching techniques is possible with careful preprocessing and similarity computation. Going forward, we plan to study the problem of mixed mineral identification using more single-mineral data from the new PDS implementation of RRUFF, as well as spectra of mineral mixtures created in our own laboratory. We believe that the accuracy of mineral identification from mixtures is fundamentally dependent on the success of single-phase recognition.

Acknowledgements

This work was supported by NSF grant DUE-1140312 and NASA grant NNA14AB04A to the RIS^{4}E node of the Remote, In Situ, and Synchrotron Studies for Science and Exploration (SSERVI). This is SSERVI paper #2014-234. We thank Bob Downs for making the RRUFF data available to us in a convenient format.

References

[1] W. G Kong, A. Wang, Planetary laser Raman spectroscopy for surface exploration on C/D-type asteroids—a case study, in *Lunar and Planetary Science Conference*, vol. 1, Lunar and Planetary Institute, Houston, TX, **2010**, pp. 2730.
[2] S. Michael Angel, R. G Nathaniel, Shiv K.S, Chris M.K, *Appl. Spectros.* **2012**, *66*, 137–150.
[3] P. Sobron, C. Lefebvre, A. Koujelev, A. Wang, Why Raman and LIBS for exploring icy moons? in *Lunar and Planetary Institute Science Conference Abstracts*, vol. 44, Lunar and Planetary Institute, Houston, TX, **2013**, pp. 2381.
[4] S. K Sharma, P. G Lucey, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2003**, *59*, 2391–2407.
[5] A. Wang, L. A. Haskin, A. L. Lane, T. J. Wdowiak, S. W. Squyres, R. J. Wilson, L. E. Hovland, K. S. Manatt, N. Raouf, C. D. Smith, *Journal of Geophysical Research: Planets* **2003**, *108*, 1991–2012.
[6] P. Sobron, F. Sobron, A. Sanz, F. Rull, *Appl. Spectros.* **2008**, *62*, 364–370.
[7] Z. C. Ling, A. Wang, B. L. Jolliff, C. Li, J. Liu, W. Bian, X. Ren, Y Su, Raman spectroscopic study of quartz in lunar soils from Apollo 14 and 15 missions, in *Lunar and Planetary Science Conference*, vol. 40, Lunar and Planetary Institute, Houston, TX, **2009**, pp. 1823.
[8] J. L. Lambert, J. Morookian, T. Roberts, J. Polk, S. Smrekar, S. M. Clegg, R. C. Weins, M. D. Dyar, A. Treiman, Standoff LIBS and Raman spectroscopy under Venus conditions, in *Lunar and Planetary Science Conference*, vol. 41, Lunar and Planetary Institute, Houston, TX, **2010**, pp. 2608.
[9] F. Rull, A. Sansano, E. Díaz, C. P. Canora, A. G. Moral, C. Tato, M. Colombo, T. Belenguer, M. Fernández, J. A. R. Manfredi, R. Canchal, B. Dávila, A. Jiménez, P. Gallego, S. Ibarmia, J. A. R. Prieto, A. Santiago, J. Pla, G. Ramos, C. Díaz, C. González, ExoMars Raman laser spectrometer for ExoMars, in *SPIE Optical Engineering + Applications*, vol. 8152, Society of Photo-Optical Instrumentation Engineers (SPIE), San Diego, CA, **2011**, pp. 81520J–81520J. International Society for Optics and Photonics.
[10] F. Rull, S. Maurice, E. Diaz, C. Tato, A. Pacros, Rls Team, The Raman laser spectrometer (RLS) on the ExoMars 2018 rover mission, in *Lunar and Planetary Institute Science Conference Abstracts*, vol. 42, Lunar and Planetary Institute, Houston, TX, **2011**, pp. 2400.
[11] Jet Propulsion Laboratory, SHERLOC to micro-map Mars minerals and carbon rings, **2014**. [Online; accessed 24-September-2014].
[12] R. T. Downs, The RRUFF Project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals, in *Program and Abstracts of the 19th General Meeting of the International Mineralogical Association in Kobe, Japan*, vol. 1, International Mineralogical Association, Kobe, Japan, **2006**, pp. O03–13.
[13] T. Laetsch, R. T. Downs, in *19th General Meeting of the International Mineralogical Association, Kobe, Japan*, vol. 1, International Mineralogical Association, Kobe, Japan, **2006**, pp. 23–28.
[14] M. M. Paradkar, J. Irudayaraj, *Food Chem.* **2002**, *76*, 231–239.
[15] M. S. Gilmore, B. Bornstein, M. D. Merrill, R. Castaño, J. P. Greenwood, *Icarus* **2008**, *195*, 169–183.
[16] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, L. M. C. Buydens, *Chemom. Intell. Lab. Syst.* **2004**, *73*, 169–179.
[17] M. Gallagher, P. Deacon, Neural networks and the classification of mineralogical samples using X-ray spectra, in *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, vol. 5, IEEE, Singapore, **2002**, pp. 2683–2687.
[18] G. Lopez-Reyes, P. Sobron, C. Lefebvre, F. Rull, *Am. Mineral.* **2014**, *99*, 1570–1579.
[19] T. L. Roush, R. Hogan, Automated classification of visible and near-infrared spectra using self-organizing maps, in *Aerospace Conference, 2007, IEEE*, vol. 1, Institute of Electrical and Electronics Engineers (IEEE), Big Sky, MT, **2007**, pp. 1–10.
[20] S. Bayraktar, B. Labitzke, J. Bader, R. Bornemann, P. Haring Bolivar, A. Kolb, Efficient, robust, and scale-invariant decomposition of Raman spectra, in *Signal and Image Processing Applications (ICSIPA), 2013, IEEE International Conference on*, vol. 1, IEEE, Melaka, Malaysia, **2013**, pp. 317–321.
[21] S. Lowry, D. Wieboldt, D. Dalrymple, R. Jasinevicius, R. T. Downs, *Spectroscopy* **2009**, *24*, 1–7.
[22] S. T. Ishikawa, V. C. Gulick, *Comput. Geosci.* **2013**, *54*, 259–268.
[23] J. A. Jaszczak, *Rocks & Minerals* **2013**, *88*, 184–189.
[24] K. Carron, R. Cox, *Anal. Chem.* **2010**, *82*, 3419–3425.
[25] V. Baeten, P. Hourant, M. T. Morales, R. Aparicio, *J. Agric. Food Chem.* **1998**, *46*, 2638–2646.
[26] G. Lopez-Reyes, F. Rull, G. Venegas, F. Westall, F. Foucher, N. Bost, A. Sanz, A. Catalá-Espí, A. Vegas, I. Hermosilla, A. Sansano, J. Medina, *Eur. J. Mineral.* **2013**, *25*, 721–733.
[27] W. Schumacher, M. Kühnert, P. Rösch, J. Popp, *J. Raman Spectrosc.* **2011**, *42*, 383–392.
[28] N. V. Vagenas, C. G. Kontoyannis, *Vib. Spectros.* **2003**, *32*, 261–264.
[29] L. A. Haskin, A. Wang, K. M. Rockow, B. L. Jolliff, R. L. Korotev, K. M. Viskupic, *J. Geophys. Res-Planets (1991–2012)* **1997**, *102*, 19293–19306.
[30] R. Perez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *J. Raman Spectrosc.* **2004**, *35*, 808–812.
[31] E. Kriesten, F. Alsmeyer, A. Bardow, W. Marquardt, *Chemom. Intell. Lab. Syst.* **2008**, *91*, 181–193.
[32] I. H. Rodriguez, G. Lopez-Reyes, D. R. Llanos, F. Rull Perez, Automatic Raman spectra processing for Exomars, in *Mathematics of Planet Earth*, vol. 1, Springer Berlin Heidelberg, Berlin, Germany, **2014**, pp. 127–130.
[33] R. V. Gaines, J. D. Dana, E. S. Dana, *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, Wiley, Hoboken, NJ, **1997**.
[34] Inc., Spectrum Square Associates. RAZOR LIBRARY: resolution enhancement, smoothing, derivatives, peak picking, peak fitting, and baseline estimation using Bayesian, maximum likelihood, and maximum entropy spectral analysis methods, **1998**. [Online; accessed 29-September-2014].
[35] C. J. Cobas, M. A. Bernstein, M. Martin-Pastor, P. G. Tahoces, *J. Magn. Reson.* **2006**, *183*, 145–151.
[36] W. Dietrich, C. H. Rüdel, *J. Magn. Reson. (1969)* **1991**, *91*, 1–11.
[37] P. H. C. Eilers, H. F. M. Boelens, Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, Leiden University, Leiden, NL, **2005**.
[38] Z. M. Zhang, S. Chen, Y. Z. Liang, *Analyst* **2010**, *135*, 1138–1146.
[39] G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, M. W. Blades, *Appl. Spectrosc.* **2005**, *59*, 545–574.
[40] K. E. Kuebler, B. L. Jolliff, A. Wang, L. A. Haskin, *Geochim. Cosmochim. Acta* **2006**, *70*, 6201–6222.
[41] I. T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, Springer-Verlag, Berlin, Germany, **2002**.
[42] A. Savitzky, M. J. E. Golay, *Anal. Chem.* **1964**, *36*, 1627–1639.
[43] M. M. Deza, E. Deza, *Encyclopedia of Distances*, Springer, Berlin, Germany, **2009**.
[44] H. Sakoe, C. Seibi, *Acoustics, Speech and Signal Processing, IEEE Transactions on* **1978**, *26*, 43–49.
[45] M. Vlachos, G. Kollios, G. Dimitrios, Discovering similar multidimensional trajectories, in *Data Engineering, 2002. Proceedings. 18th International Conference on*, IEEE, San Jose, CA, **2002**, pp. 673–684.

[46] T. Eiter, H. Mannila, Computing discrete Fréchet distance. *See Also*, **1994**.
[47] G. Tomasi, F. van den Berg, C. Andersson, *J. Chemom.* **2004**, *18*, 231–241.
[48] P. H. C. Eilers, *Anal. Chem.* **2004**, *76*, 404–411.
[49] H. J. Ramaker, E. N. M. van Sprang, J. A. Westerhuis, A. K. Smilde, *Anal. Chim. Acta* **2003**, *498*, 133–153.

