MACHINE LEARNING TOOLS FOR MINERAL RECOGNITION AND CLASSIFICATION FROM RAMAN SPECTROSCOPY. C. Carey[1], T. Boucher[1], S. Mahadevan[1], M. D. Dyar[2], and P. Bartholomew[3], [1]School of Computer Science, University of Massachusetts at Amherst, Amherst MA 01003, ccarey@cs.umass.edu, boucher@cs.umass.edu, mahadeva@cs.umass.edu, and [2]Department of Astronomy, Mount Holyoke College, South Hadley, MA 01075 mdyar@mtholyoke.edu, [3]Dept. of Biology and Environmental Sci., University of New Haven, West Haven, CT 06516, PBartholomew@newhaven.edu.

Introduction: Tools for materials identification based on Raman spectroscopy 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 tested by us has several limitations, including:

1. 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.

2. 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.

3. Existing search/match software is computationally cumbersome, resulting in long run times when the reference database is large.

Further complications are introduced by different styles and shapes of peaks, as well as degrees of photoluminescent interference.

For this project, we seek to evaluate full-spectrum matching algorithms through the application of modern machine learning methods, which show great promise in dealing with the difficulties noted above.

Spectra Used: For this project, we used a subset of spectra from the RRUFF database [1]. Only data from unoriented samples collected at random orientations were used. Spectra had been processed (background-removed), and instrumental artifacts were removed. Unprocessed spectra with overwhelming specimen fluorescence were excluded. Each mineral name was matched with its four-part Dana classification number in which the first number is mineral class, the second is mineral type, the third is the mineral group, and the fourth is the specific mineral species.

Techniques: *Vector similarity metrics*. The most popular form of full-spectrum matching is comparison under a vector similarity metric. Of these, cosine similarity is both chemically relevant and used in the popular CrystalSleuth software [1]. In our experiments, all spectra were cropped and resampled to a common set of bands, resulting in a 1715-dimensional vector representation. High similarity is achieved when two spectra have many matching bands with high intensity. Bands in which either target or query spectra have low intensity do not contribute significantly to the overall similarity score.

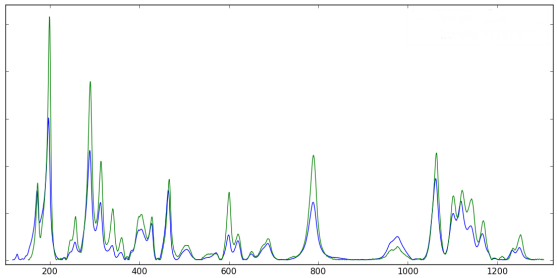

Figure 1. 532 nm laser Raman spectra of trolleite samples 30565 and 32267 from the RRUFF database showing variations in peak intensities that occur even in samples of the same mineral species and laser energy.

Preprocessing. Due to differences in laser polarization and sample crystal orientation, peaks in matching spectra of the same species often vary in intensity (**Figure 1**), inducing dissimilarity under the cosine metric. To counteract this issue, we apply several steps of nonlinear, monotonic preprocessing, including square root squashing, maximum value normalization (rescaling), and sigmoid squashing. Each step preserves relative ordering while mitigating the effect of peak intensity differences. This combination of preprocessing steps increases match accuracy. Other approaches, such as smoothing and Savitzky-Golay derivatives, did not increase match accuracy in our tests.

Dimension reduction. With high dimensional data, methods for reducing the dimensionality of the matched vectors can be effectively employed. One common method is principal components analysis (PCA), which solves for a linear mapping from the input space to a lower-dimensional feature space while preserving as much variance as possible. Although PCA preprocessing was shown to improve accuracy in a previous study [2], our results suggest the opposite, likely because the learned linear mapping ignores locality and other chemically relevant aspects of the data.

Trajectory similarity. A downside of vector-based approaches is their requirement that all spectra be sampled in the same interval, at the same bands. Real-

Table 1. Confusion matrix of cosine similarity performance on mineral group classification, using 318 test and 52 training mineral samples (2 samples per species) from the RRUFF database. Average group accuracy is **96.5%**.

	Qtz	Ksp	Plag	Pyx	Mica	Ol	% Acc.
Qtz	22	0	0	0	0	0	100
Ksp	0	30	1	1	0	0	94
Plag	0	5	40	0	0	0	89
Pyx	0	0	2	114	1	0	97
Mica	0	0	0	1	46	0	98
Ol	0	0	0	0	0	55	100

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 wavelength-intensity pairs. One such algorithm is our novel extension of Longest Common Subsequence (LCSS) similarity [3], which uses a Minimum Bounding Envelope approach to compute matching points in a pair of trajectories.

Preliminary Results: We compare against the results published by Ishikawa and Gulick [2] on a subset of minerals from the RRUFF data set; they obtained a prediction accuracy of 80.4%. Our average accuracies resulting from use of the preprocessing steps noted above, in tandem with a cosine similarity-based first nearest neighbor classifier, are given in **Table 1**. These results demonstrate that simple full-spectrum matching techniques can achieve equivalent or better accuracy than sophisticated PCA and neural-network classifiers.

In **Figure 2**, we examine the relative pairwise cosine similarity between samples from the same RRUFF mineral subset. Each square represents pairwise similarity between all 370 samples, which have been ordered such that samples of the same group and species are adjacent. Comparing the ground-truth similarities from the top image to the experimental cosine similarities in the bottom image shows that most mineral groups are distinctly separated by the metric.

In addition to the cosine similarity algorithm, we evaluate the new LCSS-based trajectory matching technique, with results on the RRUFF subset reported in **Table 2**. The intensity values were scaled using the same preprocessing steps from the cosine similarity

Table 2. Confusion matrix of trajectory-based similarity performance on mineral group classification using RRUFF samples from Table 1. Average group accuracy is **93.1%**.

	Qtz	Ksp	Plag	Pyx	Mica	Ol	% Acc.
Qtz	22	0	0	0	0	0	100
Ksp	0	29	3	0	0	0	91
Plag	0	3	42	0	0	0	93
Pyx	0	0	2	104	11	0	89
Mica	0	0	0	3	44	0	94
Ol	0	0	0	0	0	55	100

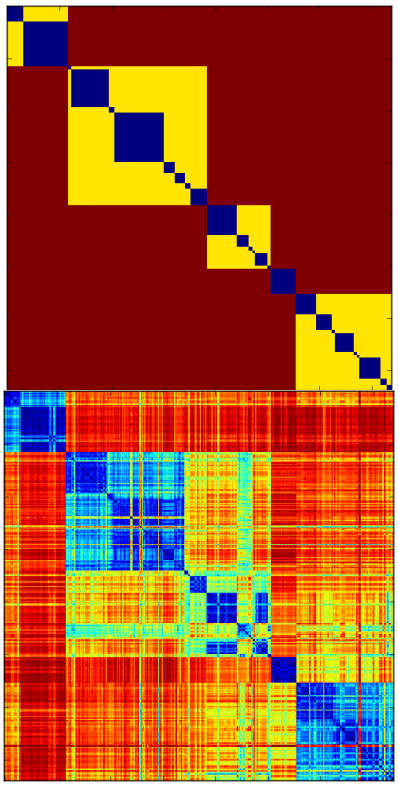

Figure 2. Intraclass similarity of minerals used in Table 1. (top) Ground truth, in which yellow denotes a group-level match, and blue denotes species-level match. (bottom) Cosine similarity, from red (low) to blue (high).

experiment, but no resampling or cropping was required. This approach achieves similar performance to the cosine similarity algorithm, while providing maximum flexibility regarding the length and sampling density of each spectrum.

Conclusions: 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. These techniques demonstrate the potential for expanding the application of full-spectrum algorithms to larger data sets and other kinds of materials.

Acknowledgement: This work was supported by NSF grant DUE-1140312 and NASA grant NNA14AB04A to the Remote, In Situ, and Synchrotron Studies for Science and Exploration SSERVI. We thank Bob Downs for making the RRUFF data available to us in a convenient format.

References: [1] Downs, R.T. (2006) 19th *Internat. Mineralog. Assoc. Meeting*, Kobe, Japan, 003-13. [2] Ishikawa, S.T. et al. (2013) *Computers & Geosciences, 54,* 259-268. [3] Vlachos, M., et al. (2002) *Data Engineering, IEEE*. 673-684.