IMPROVING MATCHING ACCURACY IN RAMAN SPECTROSCOPY BY QUANTIFYING THE WAVENUMBER SHIFT IN RAMAN SPECTROSCOPY BETWEEN INSTUMENTS. T. Mullen[1], M. D. Dyar[2], M. Parente[1], L. Breitenfeld[2], [1]Dept. Electrical & Computer Engineering, Univ. of Massachusetts, Amherst MA 01003 (thmullen@umail.edu), [2]Mount Holyoke College, Dept. of Astronomy, South Hadley, MA 01075.

Introduction:. As technology for *in situ* planetary exploration evolves, Raman spectroscopy will likely become an important, ubiquitous tool. Not only can Raman identify mineralogy (not unlike x-ray diffraction), but it can also be used to fingerprint important organic molecules significant for explorations involving the origins of life. The *ExoMars* Raman-LIBS and *Mars 2020* SuperCam (both 532 nm lasers) and SHERLOC instruments (248.6 nm) all employ Raman spectrometers to probe Mars surface targets and derive information about mineralogy and organics. Interpreting these data may be challenging because there are almost no mineral data collected with a deep-UV laser, few data for mineral powders, and virtually no data on mixtures except those produced in our group [1-5]. A preliminary cross-instrument comparison [3] (Figure 1) shows large variations among Raman instruments (noise, peak position, relative peak intensity), requiring new techniques to mitigate the impact that variations in instruments have on match success.

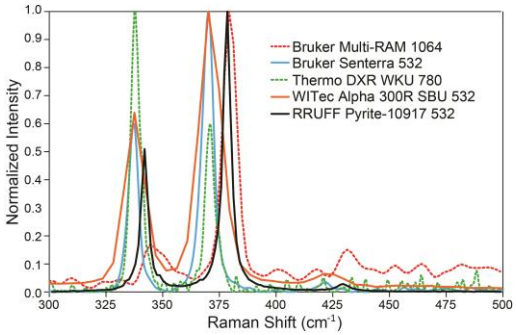

Figure 1. Close-up of normalized Raman spectra of the same pyrite sample collected with different instruments and laser energies for an interlaboratory comparison (Dyar et al. 2016). Pyrite data from the RRUFF database (different sample) shown in black.

Background: Several workers have considered the issues involved with automatic matching and identification of Raman and other types of spectra. Early efforts relied on expert knowledge of spectral features, but more recent approaches have used a wide range of statistical and machine learning tools including support vector machines [6], artificial neural networks [7,8], similiarity based methods [9], and full-spectrum matching[10, 11]. Most methods employ some combination of spectrum pre-processing steps to reduce the influence of noise and fluorescence [12, 13], and some also project spectra into a lower-dimensional feature space, typically using Principal Components Analysis [14,15] but *they focus on pure phase identification*. For mixtures, [16] uses point counting for mineral identification, but its use is restricted to applications where the Raman beam size is similar to or smaller than individual mineral grains. Models for automated identification of minerals using univariate analysis [8, 17-19] are not fully adaptable to mineral mixtures. In all these cases, mineral identification was based on relatively small, single-source reference libraries.

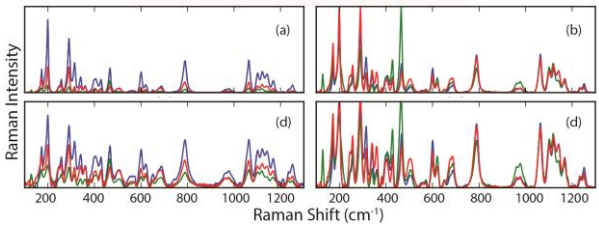

Figure 2. Visualization of the effects of various preprocessing steps from Carey et al. [20]. (a) Three resampled, unprocessed spectra of trolleite. (b) The same data, rescaled by normalizing to max-imum value. (c) Data rescaled using the square root of each wavelength's intensity. (d) A combination of (b) and (c), performing square root squashing, then maxi-mum value normalization, then sigmoid squashing. The last step scales intensity using the sigmoid function.

Optimal identification of minerals on Mars will require use of multiple spectral libraries to leverage all publicly-available sources of reference information. To create a master database for broad and flexible use in this field, we need a method to align spectra from disparate sources – i.e., make spectra of the same mineral look the same, even when they come from different instruments. However, differences in sample crystal orientation, grain size of powders, laser polarization, focus, and other instrumental parameters can have major effects on spectra, even on the same samples and identical instruments. To ameliorate these differences and allow unification of disparate data, [20] demonstrated the success of careful tuning of preprocessing steps (c.f., **Figure 2**) such as baseline removal, squashing, smoothing, and the order in which these are undertaken using a mathematical generalization. Our group is currently pursuing reinforment learning approaches to mitigate these variations, all of which affect peak *intensity* in Raman spectra. However, these techniques also require dealing with energy calibrations of data from different instruments.

Methods: Here we compare the location of strong peaks *of the same sample* taken using the Bruker MultiRAM 1064 and the Caltech Renishaw 783 instruments with the peaks found in spectra taken with Bruker Senterra 532. Data were resampled from 300 cm^{-1} to 1500 cm^{-1} to have matching resolution, $2cm^{-1}$ for comparing the MultiRAM and Senterra and $1cm^{-1}$ for comparing the Renishaw and Senterra. As a final preprocessing step, we used a Savitzky Golay filter to smooth out some of the noise.

Next, we found the five largest peaks using "detect_peaks" [21] to identify peaks separately in both spectra. We checked that the peaks in spectra from instrument 1 had a corresponding peak in spectra from instrument 2. Because *we were only comparing spectra from the same exact sample*, they should have the same peaks at the same wavenumber. If there was a significant overlap, then we labeled a peak as "strong" and not due to noise. Using these peaks, we quantified the shift between instruments.

Raman peaks are often modeled as Voigt functions, thus, the extracted peaks were fit using the Voigt function [22] to extract the center of the Voigts and find the wavenumber shift. This provided a measure of confidence that we were using the true center of the peak. By taking the difference between the corresponding peaks from the same sample in spectra taken by different instruments, we could assessed the magnitude of wavenumber shifts among different instruments. We followed the process to find the individual peak shifts in **Figure 3**.

Results: The Bruker Senterra 532 and the Caltech Renishaw 783 are well aligned (Figure 3, top). All of the compared peaks had shifts that were within the the resampling resolution of $1cm^{-1}$.

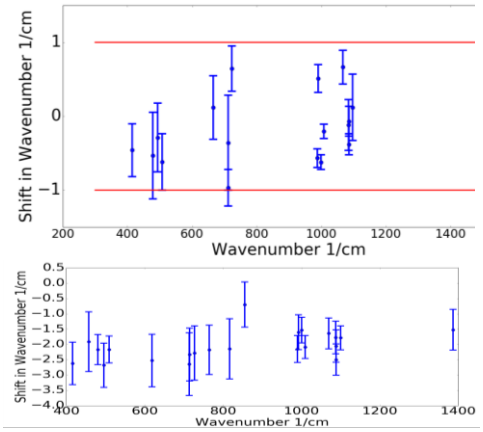

Figure 3. Each point represents the shift between corresponding peaks from the same sample. We compare Caltech Ren-ishaw 783(top) and the Bruker MultiRAM 1064 (bottom) to the Bruker Senterra 532. The shifts fall within the resolution of the resolution of the resampling in the Caltech, but not in the MultiRAM.

The Bruker MultiRAM 1064 data did not align well with the Bruker Senterra 532 (Figure 3, bottom). There was a consistent negative shift through all wavenumbers that varied in a nonlinear fashion. The shifts in wavenumber are greater than the resampling resolution and can be seen clearly individual spectra (**Figure 4**). If the wavenumber shifts are not taken into account they will cause significant difficulties in the Raman matching.

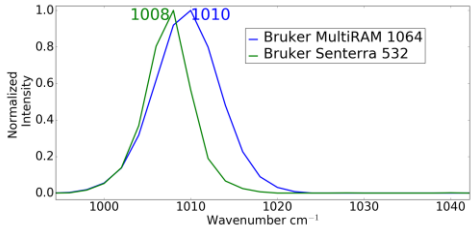

Figure 4. Spectra from sample Gypsum 34 taken with the Bruker MultiRAM 1064 and the Bruker Senterra 532. There is an clear shift even though the two spectra are from the same sample.

Discussion: Previous work demonstrated that matching accuracy against a spectral library improved significantly when pre-processing techniques were applied to peak intensities (i.e., counts displayed on the *y* axis) for Raman data [20]. Here we show that attention must also be paid to aligning wavenumber (the *x* axis) when combining data from disparate instruments and data sets. Only through careful alignment of both intensities and energy can spectral libraries be effectively used for matching algorithms.

References: [1] Breitenfeld, L. et al. (2016) *LPS XLVII*, Abstract #2430. [2] Breitenfeld, L. et al. (2016) *LPS XLVII*, Abstract #2186. [3] Dyar, M. D. et al. (2016) *LPS XLVII*, Abstract #2240. [4] Berlanga, G. et al. (2017) *LPS XLVIII*, Abstract #2255. [5] Clegg, S.M. et al. (2014) *Appl. Spectros.*, 68, 925-936. [6] Thissen, U. et al. (2004) *Chemon. Intel. Lab. Syst*, *73*, 169-179. [7] Gallagher, M. and Deacon, P. (2002) *ICONNIP*, *5*, 2683-2687. [8] Lopez-Reyes, G. et al. (2014) *Amer. Mineral*, *25*, 721-733. [9] Sobron, P. et al. (2008) *Appl. Spectrosc*, *62*, 364-370. [10] Bayrakaktar, S. et al. (2013) *ICSIPA*, 317-321. [11] Lowry, S. et al. (2009) *Spectrosc.*, *24*, 1-7. [12] Jaszczak, J. A. (2013) *Rocks Mins*, *88*, 184-189. [13] Carron, K. and Cox, R. (2010) *Anal. Chem.*, *82*, 3419-3425. [14] Baeten, V. et al. (1998) *J.Agrig. Food Chem.*, *46*, 2638-2646. [15] Ishikawa, S. T. and Gulick, V.C. (2013) *Comp. Geosci*, *54*, 259-268. [16] Haskin, L.A. et al. (1997) *J. Geophys. Res. Planets*, *102(E8)*, 19293-19306. [17] Perez-Pueyo, R. et al. (2004) *J. Raman Spectrosc. 35*, 808-812. [18] Kriesten, E. et al. (2008) *Chemom. Intell. Lab. Syst.*, *91*, 121-193. [19] Rodriguez, I. H. et al. (2014) *Mathematics of Planet Earth*, 127-130. [20] Carey, C. et al. (2015) *J. Raman Spectros.* 894-903. [21] Duarte, M. (2015) https://github.com/demotu/BMC. [22] Dong, X. and Dai, L. (2012) *Asain Journal of Chemistry*, 4257–4262.