This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Predicting olivine composition using Raman spectroscopy through band shift and multivariate analyses

LAURA B. BREITENFELD,[1] M. DARBY DYAR,[1] CJ CAREY,[2] THOMAS J. TAGUE, JR.,[3] PENG WANG[3], TERRY MULLEN[4], AND MARIO PARENTE[4]

[1]Department of Astronomy, Mount Holyoke College, South Hadley, MA 01075, U.S.A.

[2]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst MA 01003, U.S.A.

[3]Bruker Optics, Inc., Billerica, MA 01821, U.S.A.

[4]Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst MA 01003, U.S.A.

Revision #2

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

ABSTRACT

Olivine group minerals are ubiquitous in extrusive igneous rocks, and play an important role in constraining equilibria for samples in the upper mantle and above. All Raman spectra of the olivine group minerals in the solid solution between forsterite (Fo, Mg_2SiO_4) and fayalite (Fa, Fe_2SiO_4) have a high intensity doublet between 800 and 880 cm^{-1}. Previous studies used small sample suites with limited compositional ranges and varying spectrometers to relate energy shifts of these two bands to Mg/Fe contents. In this work, Raman spectra of 93 olivine samples were acquired on either Bruker's 532 nm (laser wavelength) Senterra or BRAVO (785/852.3 nm) spectrometer. This paper compares the two-peak band shift univariate method with two multivariate methods: partial least squares (PLS) and the least absolute shrinkage operator (Lasso). Datasets from several instruments are also examined to assess the most accurate method for predicting olivine composition from a Raman spectrum.

Our 181-spectra PLS model is recommended for use when determining olivine composition from a Raman spectrum. For Raman spectra of mixed phases where only the olivine doublet can be identified, composition can best be determined using the position of the peak ca. 838-857 cm^{-1} through use of the equation $\%Fo = -0.179625x^2 + 310.077x - 133717$ regression equation (where x = DB2 centroid in units of cm^{-1}).

In situ methods for predicting mineral composition on planetary surfaces are critically important to extraterrestrial exploration going forward; of these, Raman spectroscopy is likely the best, as evidence by the impending deployment of several Raman instruments to Mars (*ExoMars* and *Mars 2020*). More broadly, application of machine learning methods to spectral data processing have implications to multiple fields that use spectroscopic data.

Keywords: Raman spectroscopy, olivine, forsterite, fayalite, PLS, Lasso

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

INTRODUCTION

Olivine group minerals control many of the properties of Earth's upper mantle, affect rheology, and may be diagnostic of crystallization temperature in terrestrial and extraterrestrial rocks. Their solid solution commonly spans the range between forsterite (Fo, or Mg_2SiO_4) and fayalite (Fa, or Fe_2SiO_4) with minor substitutions of alternative cations such as Mn and Ni. Because olivine composition provides an important petrogenetic indicator, development of convenient methods to measure it without microanalytical techniques that require sample preparation are desirable. This paper focuses on determination of olivine composition using Raman spectroscopy for this purpose. It has the potential to enable compositions to be conveniently determined in the laboratory, in field identifications with portable units, and on other planets such as Mars.

This problem has been extensively studied using conventional regression-based approaches, but generalization of their results is arguable given their very small (<20) sample suites and coverage of olivine composition. These prior studies (Kuebler et al. 2006, Foster et al. 2007, Gaisler and Kolesov 2007, Mouri and Enami, 2008, Yasuzuka et al. 2009, Ishibashi et al. 2011) have related olivine composition to the peak positions of a high intensity doublet in the range of 800-880 cm^{-1} (Figure 1). Peak centroids are regressed against composition to obtain an equation suitable for prediction of composition in unseen samples. Previous workers have used R^2 values to characterize their prediction algorithms, making their accuracy difficult to quantify and comparisons across models impossible. Moreover, a variety of Raman instruments with varying resolution and laser wavelength, and a combinations of single crystal and powdered samples, were used. It is thus difficult to assess which model to use to estimate olivine composition on unseen data from a different spectrometer than the one used in each study.

In this study, all known publicly-accessible olivine Raman data are considered. In addition, new Raman spectra were acquired from a suite of 93 well-characterized synthetic and naturally-occurring olivines using Bruker BRAVO and Senterra spectrometers. The accuracies of linear regression (univariate) methods for various datasets are quantified to enable useful comparisons. Univariate methods are compared and contrasted with two multivariate analysis approaches: partial least squares (Stone and Brooks 1990) and the least absolute shrinkage operator (Hastie et al. 2009) to evaluate the best prediction models for use with Raman spectra of olivines to determine composition with known accuracy.

BACKGROUND

Assignments of Raman active forsterite and fayalite modes have evolved over time (Table 1). Generally, forsterite Raman bands above 500 cm^{-1} can be classified as internal movement within the $(SiO_4)^{4-}$ tetrahedra. Below this threshold energy, peaks are caused by rotation and translation of the tetrahedra as well as divalent cation motion. Forsterite and fayalite have 84 vibrational modes; only 36 are Raman active ($11A_g + 11\ B_{1g} + 7B_{2g} + 7B_{3g}$) (Mckeown et al. 2010).

Key to this study are the two principle Raman bands that form a doublet composed of five vibrational modes ($2A_g + 2B_{1g} + B_{2g}$) (Table 1). This doublet occurs between ~815-825 cm^{-1} (DB1) and ~838-857 cm^{-1} (DB2) (Kuebler et al. 2006) and is primarily attributed to A_g, though B_{1g} and B_{2g} also affects the shape and intensity of the spectrum. The energy shift of the A_g stretch from the SiO_4 tetrahedra is caused by changes in site geometry due to cation substitutions in adjacent sites. Cation substitutions between forsterite and fayalite thus result in band shifts (Figure 2) as well as changes in the shape and intensity of the peaks. Many previous workers (Table 2) have used the peak centroids of the DB1 and DB2 doublet peaks to derive olivine

composition. However, this practice does not allow other information in the spectra to be utilized, such as shifts arising from minor modes that affect the shape of the primary doublet and give rise to other, more subtle features elsewhere in the wavenumber range.

Other peaks within olivine spectra have been utilized rarely for prediction of composition, such as ~200-230 cm^{-1}, ~290-310 cm^{-1}, ~410-440 cm^{-1}, ~540-553 cm^{-1}, ~881-883 cm^{-1}, ~914-920 cm^{-1}, 950-966 cm^{-1} (Table 2). However, these features are relatively low in intensity compared to those of the DB1 and DB2 doublet, making fitting of peak centroids difficult and less accurate. Raman bands caused by different vibrational modes should not be affected by octahedral substitutions. For example, features between 400 and 700 cm^{-1} have been attributed to the internal bending modes of the anion, which have minimal centroid shifts (Chopelas, 1991; Kuebler et al., 2006).

This study evaluates the relative usefulness of the most prominent bands in the Raman spectra of olivine group minerals using a combination of conventional regression/peak fitting methods and more recently-developed multivariate methods. The latter have the advantage of weighing the relative importance of different spectral energies in determining Fe/Mg ratio, enhancing our understanding of the underlying physical processes that give rise to the features.

METHODS

Sample provenance

Natural samples (Table 3) came from collections of the Mineral Spectroscopy Lab at Mount Holyoke, the National Museum of Natural History (NMNH, Smithsonian), and from S.A. Morse (University of Massachusetts Amherst) (Morse 2001). This is the largest suite of naturally occurring olivine samples studied by Raman (or any other type of) spectroscopy. Roughly one-third of the natural samples came from previous studies of olivines from mantle xenoliths

(McGuire et al. 1991, Dyar et al. 1989, Dyar et al. 1992) or Fe^{3+}-bearing samples studied by Schaefer (1983), Banfield et al. (1992), and Dyar et al. (1998). Another group of samples was provided by S.A. Morse of the University of Massachusetts Amherst. They come from the Kiglapait layered mafic body, a large 1.3 Ga layered intrusion on the coast of Labrador, Canada (Morse 1996, Morse 2001). As the original melt crystalized, the Fe/Mg ratio of the remaining liquid changed, so a range of olivine compositions were produced. Lower Mg and higher Fe contents occurred successively higher within the intrusion. Finally, several samples came from the NMNH (see Table 3).

Naturally occurring olivine typically has high Fo content of roughly 89.5%. For a solid %Fo prediction model, wide representation of the Fo-Fa continuum is needed. Because samples with intermediate Fo/Fa content are rare in nature (except at specific localities such as the Kiglapait, as noted above), synthetic samples were added to our collection of naturally formed olivines to represent %Fo from 0 to 100 (see Dyar et al. 2009 for sample descriptions). Synthetic samples were synthesized by Donald Lindsley in his laboratory at SUNY Stony Brook. First, a silicon and hematite mixture was ground for 1-2 hours with ethanol. Next, an iron sponge was added and for less than one hour grinding continued. The product was enveloped in silver foil and put in a glass silicon capsule. The center of the capsule was drawn out into a capillary while one end of the capsule was sealed, leaving the sample by the sealed end. Near the open end of the capsule, and Fe getter was placed. For 10-20 minutes, the capsule was placed into a ~800°C vertical tube furnace (the Fe getter remained at ~600°C). Finally, the capsule was removed from the furnace and sealed across the capillary. The completely sealed capsule section, which contained the sample, was next cooked for 10 days in a horizontal tube furnace at ~920-940°C (Sklute, 2006). This sample suite has been studied with a wide range of other spectroscopic

techniques (Dyar et al. 2009, Lane et al. 2011, Isaacson et al. 2014).

Sample characterization

Olivines examined were either a single crystal or powdered samples. To produce a powdered sample, each sample was first visually inspected and handpicked for purity. Then each grain was treated using oxalic acid (2 tsp. in 2 gal. of water) for one hour to remove surface weathering, followed by three cycles of washing and rinsing with clean water. As needed, samples were either crushed in a tungsten shatterbox or ground by hand in a diamonite mortar. Because crystal orientation affects the Raman spectrum, we chose to study both single crystals and powders (Price et al. 1987), affording the opportunity to compare those results. The spot sizes of the Senterra and BRAVO spectrometers differ, so that single crystals were analyzed with the Senterra while the BRAVO examined powders. However, repeated analyses of the same sample on each instrument showed no evidence for heterogeneity, as expected given our careful sample preparation and use of homogeneous starting material. The only difference was a consistent offset (as discussed below) due to differences in calibration.

Many natural samples from Dyar's collections already had published compositions (Table 3) that included Mössbauer studies to determine Fe^{3+} contents. Where needed, additional samples were analyzed by Mössbauer spectroscopy using standard methods (Sklute, 2006). Rh was used on a WEB Research Co. model W100 spectrometer equipped with a Janus closed-cycle He refrigerator. Run times ranged from 2-12 hours; results were calibrated against α-Fe foil. Typical count rates were between 500,000 and 900,000 non-resonant counts/hour. Most samples contained no Fe^{3+}.

As needed, new electron microprobe analyses of 10 spots on each sample were acquired either by Molly McCanta at the University of Tennessee in Knoxville or at Brown University by

Joseph Boesenberg; in both cases, standard operation conditions were used. Figure 3 shows the calculated %Fo for each sample that was determined by normalizing the contents to contain only Mg and Fe, as commonly done with the formula $\%Fo = (100\times Mg)/(Mg+Fe_{total})$, where $Fe_{total}=\Sigma Fe^{2+}+Fe^{3+}$. This represented only a minor adjustment because only very minor substitutions of other cations were observed, as seen in compositions of the natural samples as given in Table S1. Synthetic samples are as-named in Dyar et al. (2012a).

Raman measurements

Spectra of powdered samples were acquired on a BRAVO dual laser (785 and 852.3 nm simultaneous DuoLaser^{TM}) system (2.0 cm^{-1}/channel spectral resolution) with three sample scans and an integration time of 10s. Because the BRAVO samples a large area (~2 mm diameter), it required sample masses of >100 mg, which were only available for 25 samples. The remainder of the sample suite (68 samples) was run on a Bruker Senterra spectrometer using the 532 nm laser and a microscope attachment to probe single grains. The Senterra used 10 mW laser power for two sample scans and integrated for 10s, analyzed through a 20× objective. The highest Senterra resolution available of 0.5 cm^{-1}/channel was utilized.

The Senterra calibration was performed automatically and was anchored by the NIST standards, acetaminophen and silicon, resulting in a wavelength accuracy of 0.2 cm^{-1}. The photometric accuracy was verified using NIST traceable glass (Allen et al. 2000). The BRAVO wavelength was similarly calibrated with the wavelength accuracy being 1 cm^{-1} or better. Multiple sample scans were acquired for each sample ensuring reproducibility of the spectral data acquired. Pre-processed data that included dark subtractions and baseline removal were converted from Bruker's Opus format into ascii files and uploaded to the lab web site, currently at nemo.cs.umass/edu:54321.

Data analysis

Because many spectra showed residual features after the baseline was mostly removed by the Opus algorithm, we applied additional baseline removal using the adaptive iteratively reweighted penalized least squares (AirPLS) method (Zhang et al. 2010), which uses the sum of differences between signal and baseline to adjust weights intelligently. Smoothness is the adjustable baseline removal parameter, for which a value of 100 was used. Multiple types of normalization were tested on these data including normalizing to the maximum value (L∞norm), the sum of absolute values (L1 norm), the sum of squared values (L2 norm), and scaling to intensity at several specific energies. Normalization to the maximum peak intensity outperformed all other methods and therefore it was used in subsequent analyses throughout. Normalization was executed to account for arbitrary intensity differences between the two spectrometers. DB1 and DB2 were peak fitted for each spectrum using Gaussian and Lorentzian peak shapes and a method that simply sums all the counts in the region of interest (e.g., 800-880 cm^{-1}). Pre-processing of spectra used the superman website nemo.umass.cs.edu:54321 (Carey et al. 2017).

Next, PLS and Lasso models were applied for multivariate analysis of %Fo. PLS regresses one response variable (%Fo) against multiple explanatory variables (intensity at each channel of the spectra). PLS predictions utilize every channel of the spectral range, assigning coefficients to every single channel. Because PLS utilizes all available variables (channels) and eliminates multicollinearity (peaks whose intensities are dependent, as is the case for the doublet in the Raman spectra of olivine). This algorithm was created for the analysis of data with high collinear explanatory (p) variables, which are significantly greater in number compared to the observations (N). Therefore, $p >> N$ (Butler and Denham 2000). PLS can predict multiple

dimensional datasets and has been utilized for the specific application of spectroscopy (Wold et al. 1983). This paper utilizes PLS2 (hereafter referred to as PLS) rather than alternative versions.

Lasso is a continuous shrinkage, which allows for the production of coefficient values to be reduced even to as small as zero (Tibshirani 1995). This shrinkage is in agreement with the shrinkage parameter *t*, by shrinking the residual sum of squares based upon the sum of the absolute value of the coefficients. In other words, this method selects a subset of predictors with the strongest effect on the response variable. Unlike PLS, the Lasso produces a sparse models with few coefficients (depending on the value of the α parameter), with most channel intensities set to zero. The relative merits of PLS versus Lasso in spectroscopic methods (e.g., Dyar et al. 2012a) are just beginning to be explored and there is as yet no consensus for which method is better; their usefulness appears to be highly variable for each dataset and application-dependent.

Model comparisons

Use of the R^2 parameter to describe the fit of a regression model (here %Fo is the dependent variable and peak centroid is the independent variable) is not helpful for drawing comparisons between different models because R^2 depends on the error associated with each measurement. A more appropriate metric for cross-comparison is root mean square error (RMSE), which calculates the square root of the average difference between predicted and true %Fo. RMSE is useful in this application because it is expressed in the same unit as the measurement – in our case, %Fo. This paper uses RMSE in three different ways. Internal RMSE describes the prediction error of an expression that is created using all the data in the dataset. In other words, if there are 25 samples, the regression expression utilizes all of them. Internal RMSE is useful in comparing one model to another, but inappropriate for evaluating errors on unseen data. In contrast, leave-one-out cross-validated RMSE (LOO RMSE-CV) removes one

sample at a time, uses a regression model based on the other *n*-1 samples to predict the *n*th sample, and then repeats the process *n* times, where *n* is the number of samples in the dataset. Thus LOO RMSE-CV gives the best estimate of how the model will perform on unseen data. Finally, RMSE-test is used to describe the RMSE of comparisons between true and predicted values in completely unseen data.

UNIVARIATE (PEAK CENTROID) ANALYSIS

Univariate methods focus entirely on the two principle Raman bands in the five-mode doublet between ~815-825 cm^{-1} (DB1) and ~838-857 cm^{-1} (DB2), as discussed above and used by prior workers. Peak centroid positions (Tables S2 and S3) of the Raman spectra of our 93 synthetic and naturally-occurring olivines (Figure 2) were utilized for univariate predictions, along with data from the RRUFF database and other publications for which data were provided. There were 25 olivines for which there was sufficient sample to make measurements on the Bruker BRAVO instrument, and those 25 samples plus an additional 68 were also run on the Bruker Senterra spectrometer, which has a microbeam to enable analysis of individual grains or small clumps. Different Raman instruments can produce spectra with equivalent bands at slightly different wavenumber positions due to varying calibration protocols. Therefore, all spectra were analyzed as raw data as well as after the BRAVO dataset was aligned to the Senterra data. This was accomplished by aligning corresponding bands within spectra of 25 samples acquired on both the BRAVO and Senterra spectrometers using a method described in Mullen et al. (2018). These raw and aligned data results are compared to fits made to data taken from the RRUFF database (Table S4). Two different sets of data from RRUFF were tested: all 188 spectra of olivine group minerals, and a subset of 156 spectra designated as RRUFF*. The former group includes 32 spectra from the RRUFF site listed as “broad scan with spectral artifacts,” while the

other 156 spectra lack that designation. This annotation refers to spectra acquired over a broad energy range versus one with higher resolution. It is important to note that these data do not represent 188 different samples, but in many cases include spectra of the same samples acquired on multiple instruments, with depolarized versus polarized lasers, and on single crystals with varying orientations. Comparisons are also made to RMSE values calculated using peak positions given in papers by Kuebler et al. (2006), Yasuzuka et al. (2009) and Ishibashi et al. (2011). The equations of the first-, second-, and third-order polynomial fits are reported for the DB1 and DB2 in Table S5.

Univariate results from second-order polynomial fits to peak position versus %Fo content are summarized in Table 4, which also includes the resolution of the spectra from each dataset along with values for R^2 (coefficient of determination) of the internally cross-validated data, the internal RMSE values and LOO RMSE-CV. Linear, second-, and third-order polynomial fits relating peak centroid position to composition were created for the BRAVO and Senterra data by Breitenfeld (2017). In all cases, second-order polynomial fits to the data produced more accurate RMSE values than linear ones. Third-order polynomials produced identical or slightly better fits than second-order ones, but the improvement was negligible and not significant. Thus results in Table 4 use second-order fits, following the precedent of Kuebler et al. (2006).

While the DB1 BRAVO data conspicuously lie on a polynomial curve (Figure 4), results from the Senterra for DB1 and DB2 (both spectrometers) produce trends that are closer to linear. These differences may result from experimental parameters such as variable resolutions, excitation laser wavelength, and detector sensitivities, all of which impact the consistency of these univariate predictions.

Figure 5 displays data from sources other than this study, including the RRUFF* data, all RRUFF olivine data and results of Kuebler et al. (2006), Yasuzuka et al. (2009), and Ishibashi et al. (2011). Previous Raman studies of olivine that did not report peak centroids could not be included on this plot, such as those in Wang et al. (2004), Gaisler and Kolesov (2007), and Mouri and Enami (2008). In Figure 5, the Ishibashi et al. (2011) model predicts %Fo most accurately. It is quickly apparent that the models based on RRUFF* data perform poorly in comparison to the other univariate models, perhaps because that database combines spectra acquired on different instrument using several excitation lasers for the model.

The usefulness of applying Kuebler et al.'s (2006) models to other datasets was evaluated through the prediction of two aggregate datasets. When the Senterra + BRAVO + RRUFF* (>50%Fo) dataset is predicted using the models of Kuebler et al. (2006), the RMSE-test values for DB1 and DB2 are 9.99 and 8.10 %Fo, respectively. For the aligned Senterra + BRAVO data, the RMSE-test values for DB1 and DB2 are 19.42 and 14.64%Fo, respectively. These RMSE-test values are larger than the internal or LOO RMSE-CV values of the Kuebler et al. (2006) dataset alone (Table 4).

MULTIVARIATE (MACHINE LEARNING) ANALYSIS

In other types of spectroscopy, it has been shown that multivariate predictions using the entire spectrum produce more accurate predictions of composition. This has been demonstrated for laser-induced breakdown spectroscopy (Tucker et al. 2010, Dyar et al. 2016a) and x-ray absorption spectroscopy (Dyar et al. 2012b, Dyar et al. 2016b). Given the similarities between those data and our Raman spectra, it was expected that Raman predictions of olivine composition might follow this trend. Accordingly, both PLS and Lasso regression methods were tested on our datasets using Raman spectra acquired on the BRAVO and Senterra, along with RRUFF and

RRUFF* data for which the full spectra are available online. Because these predictions require use of the entire spectrum rather than just peak centroids, no other publicly accessible olivine Raman spectra could be included in the multivariate analyses. Model comparisons included R^2, internal PLS or Lasso RMSE, and LOO RMSE-CV, keeping in mind the caveats just discussed (Table 5).

Choice of adjustable parameter is important to the outcomes of both multivariate techniques. For PLS, the number of components in each model were tested using components ranging from 1 to 10. The value producing the lowest (best) prediction accuracy over this range was chosen (first local minimum); generally this value was 4-7 components (Table 5). Numbers of components greater than 10 might produce more accurate predictions of %Fo but they dramatically reduce the generalizability of the model to unseen data, so they were not considered. This is comparable to the concept of using very high-order polynomials to predict data – they can be quite accurate but not applicable to any other datasets.

For Lasso models, a value of α was chosen for each prediction to train the model depending on the desired "sparseness" of the model – i.e., how few channels needed to employ to predict %Fo. Variations in the value of α change the number of channels used by the model. As α increases, fewer channels are examined in the multivariate analysis prediction (Figure 6). As the number of channels in a model increases, the value of LOO RMSE-CV also decreases, showing the importance of models with a large number of channels (small α). Use of large numbers of channels may overtune the model and reduce its generalizability.

Multivariate models can be set up to use all or any subset of the spectral data/wavenumber range. As discussed above and shown in Table 2, useful wavelength ranges for prediction of olivine composition have been proposed to occur at ~200-230 cm^{-1}, ~290-310 cm^{-1},

316 ~410-440 cm^{-1}, ~540-553 cm^{-1}, ~881-883 cm^{-1}, ~914-920 cm^{-1}, and 950-966 cm^{-1}. Each of these
317 ranges was tested individually along with models covering all five of those regions as well as the
318 range from 400-700 cm^{-1} and 300-1500 cm^{-1}. Internal RMSE model accuracies are given in
319 Table 6 for BRAVO data only, Senterra data only, and then the combined datasets.
320 For the small wavelength regions taken individually, the best prediction accuracy is
321 results from the energy range between 800 and 880 cm^{-1}; although this was known from practice,
322 our data show the advantage quantitatively. The range between 400 and 700 cm^{-1}, attributed to
323 the internal bending modes of the anion as mentioned above, should cause minimal centroid
324 shifts (Chopelas, 1991; Kuebler et al., 2006), and this is reflected in the observed high RMSE
325 values for that range.
326 Interestingly, the best prediction accuracy comes from models that cover energy ranges
327 that include multiple peaks. The 300-1500 cm^{-1} PLS model resulted in RMSE values of ±3.87,
328 ±4.52, and ±5.48 %Fo for the BRAVO, Senterra, and BRAVO + Senterra models, respectively.
329 Better or comparable performance was found using only the five "useful" regions noted above:
330 ±1.67, ±4.40, and ±5.79 %Fo for PLS models and ±7.13, ±3.93, and ±5.06 %Fo for Lasso
331 models. The superiority of the five-region models over the whole-spectrum models was most
332 dramatic for the Lasso models in larger datasets.

333 **DISCUSSION**

334 **Understanding centroid variability for equivalent %Fo**

335 This study deliberately chose to include both natural and synthetic olivine samples.
336 Minor cation substitutions within the natural samples cause variations within the band centroid
337 positions affecting the accuracy of the models. Therefore samples with the same %Fo can
338 produce bands at slightly different wavenumber positions. This could be mitigated by reducing

the size of the model to only include samples that fall along the prediction line or by acquiring many spectra of the same sample to build the model. However, the goal of this study was to create a broadly applicable model to predict olivine composition in natural samples, so we chose not to remove samples that did not fall on the line of a perfect fit.

Additionally, this study intentionally acquired data on multiple spectrometers to understand the implications of making composition predictions across different laboratories. Variations observed between the two spectrometers used in this study are likely typical of comparisons that would be encountered in comparisons to data from other spectrometers (see Dyar et al. 2016c). This dilemma justifies using multivariate analysis rather than the univariate band shift method and can be mitigated by spectrometer alignment.

Restricting energy range to common compositions

The vast majority of naturally-occurring olivines contain relatively high Mg contents (Figure 3), commonly around Fo_{80-90}. So a test was developed to determine if improved accuracy could be obtained by limiting samples in the training set to those with Fo contents greater than 50%. An aggregated model of Senterra, BRAVO and RRUFF* data with %Fo values greater than 50% was constructed. The LOO RMSE-CV of this model (±4.22 for the five-region PLS model) was smaller than the original dataset (±9.45 %Fo). Model accuracy likely improves because the fayalite doublet is often poorly resolved. Therefore, in the vast majority of applications, it may only be necessary to distinguish between the compositional variations of forsterite rather than the entire olivine solid-solution. In these cases, the >50% Fo model would be preferable because of its smaller error.

Univariate analyses

Because the DB2 peak centroid covers a much wider energy shift with changing composition (compare *x*-axis limits in Figure 4 top and bottom), it might be expected that its use would result in better prediction accuracy than using the DB1 peak. This is indeed observed for data from nearly all the datasets studied (Table 4). The exceptions are in the combined BRAVO + Senterra datasets and for the datasets extracted from published papers by Kuebler et al. (2006) and Yasuzuka et al. (2009), but there is no apparent effect due to dataset size and resolution. It is notable that for the largest and most diverse datasets, DB2 fits always produce the best prediction accuracy. In particular, the model utilizing all available data, including those from the Senterra, and BRAVO instruments plus RRUFF* and other data from Kuebler et al. (2006), Yasuzuka et al. (2009), and Ishibashi et al. (2011) for which forsterite composition is >50%Fo produces the best prediction accuracy for univariate analyses, and thus its DB2 regression equation (%Fo = $-0.179625x^2 + 310.077x - 133717$) is recommended for use in applications where only peak centroids can be resolved, as might be the case in a rock spectrum where the region of interest is highly overlapped.

Multivariate analyses

It is apparent from Table 4 that both DB1 and DB2 peaks contain information about composition, so using only one of them for predictions discards useful information. In contrast, multivariate analyses offer the possibility to leverage information from anywhere in the spectra. Table 5 shows the relative accuracies of the BRAVO, Senterra, and RRUFF predictions individually and collectively in different combinations. The latter include all 188 of the RRUFF olivine data and the RRUFF*, in combinations with the BRAVO and Senterra data acquired for this project (Figure 7). As observed in the previous section, five-region models covering the

known olivine peaks show superior prediction performance over the whole-spectrum models. Why?

The answer to this question can be seen in Figure 8, which shows the magnitudes of the PLS coefficients for models covering the 300-1500 cm^{-1} range along with the channels chosen by a Lasso model with $\alpha = 0.001$ for the combined BRAVO and Senterra models. It is clear that the entire range from 300-1500 cm^{-1} is rich with information about olivine composition that has been previously unutilized in models that employ only restricted energy ranges. However, the five-region models likely outperform the whole-spectrum models because the latter may inadvertently include to unwanted features resulting from sample heating, fluorescence and, cosmic spikes. These types of noise are not consistent for each spectrum and they do not relate to Raman features resulting from compositional variations. So there is justification for using as many regions that correspond to known olivine modes as possible, and thus five-region models are used for full LOO RMSE-CV models given in Table 5.

It must be noted that use of this approach will make composition difficult to predict in practice because pure olivine is rarely encountered in field applications. In practice (as when deployed on a planetary surface), it is far more likely that the olivine will be mixed in with other phases such as glass and other minerals. Thus it is desirable to have an alternate method for predicting %Fo that isolates the olivine part of the spectra, for which the range from 800 to 880 cm^{-1} is recommended if there is no overlap from other non-olivine features.

In the largest datasets, PLS has comparable error to univariate band shift predictions, while being significantly less time consuming. Thus the most accurate %Fo prediction based on Raman spectra of pure olivine samples would be acquired at the highest resolution data possible,

pre-processed to remove baseline, normalized, and predicted using a PLS algorithm built from the maximum number of samples (here, 281).

Using solely the peak centroid to model olivine composition does not utilize information contained in other characteristics within the spectrum such as band shape, intensity, FWHM, area and anomalies/noise within the spectra. Multivariate analyses appear to overcome these effects.

PLS and Lasso models examine multiple channels within the spectra to build a %Fo prediction model. The number of coefficients per model is based on the assigned number of components from the alpha value. As the number of channels in a Lasso model increases, RMSE-CV decreases, showing the importance of models with a large number of channels. However, there is a trade-off between the generalizability of the model that is optimized by smaller numbers of channels versus improved accuracy from using larger numbers of channels.

Factors influencing prediction accuracy

Both dataset size and spectral resolution influence prediction accuracy for both univariate and multivariate models. Results presented in Table 5 inform the effects of these factors.

To test the effect of resolution, spectral data and models were resampled to 3.0 cm^{-1}/channel resolution and compared against the native resolution of each instrument, which is 2.0, 0.5 and 0.48-2.0 cm^{-1}/channel for the BRAVO, Senterra (532 nm) and RRUFF datasets, respectively. For comparison, the SuperCam instrument on Mars will have a pixel resolution of 2.5 cm^{-1} (Wiens et al. 2017), while the ExoMars RLS will use <1 cm^{-1} (Moral et al. 2018). Based on the data in this study, there is no systematic effect of spatial resolution on prediction accuracy.

The effects of dataset size can also be roughly evaluated using the data collected here. Individual datasets produce smaller (more accurate) LOO RMSE-CV values than aggregated ones because the instrument and operating conditions are identical. For example, the combined

BRAVO and Senterra datasets had PLS LOO RMSE-CV values of 7.69 for the 800-880 cm^{-1} range, while their individual, independently produced PLS LOO RMSE-CVs are 6.85 (BRAVO) and 7.06 (Senterra) (Table 5). However, these single instrument models are less generalizable. As the aggregated models increase in size and spectral diversity (i.e., instrument, laser wavelength), the LOO RMSE-CV for the multivariate models decreases and the advantages of a sole-source dataset diminish. Additionally, LOO RMSE-CV values can be reduced for aggregate datasets by aligning the spectral data of the multiple instruments (Table 5).

A five-region PLS model of the collective Senterra, BRAVO, and RRUFF* datasets with solely spectra corresponding to >50%Fo is recommended for future work. PLS (LOO RMSE-CV is 4.22 %Fo) is less time-consuming than univariate analyses (LOO RMSE-CV 4.58 %Fo), gives comparable accuracy, and is more generalizable. The PLS prediction model will aid workers using different spectrometers, incident laser wavelengths and other operating conditions. Additionally, a %Fo prediction for forsterite is more likely necessary than that of fayalite, especially for applications to Mars.

It is hoped that future workers will add Raman spectral data to the recommended models presented here, for which data are available on the lab website (nemo.cs.umass.edu:54321) (Carey et al. 2017). Increasing the number of spectra within the models on additional instruments will likely improve the accuracy of prediction results. Future workers with Raman spectra of olivines with known compositions are encouraged to contact these authors so that a new, expanded olivine prediction model can be created. Current and future improved PLS models will be available from the authors.

Parameters for model comparisons

Tabulated R^2 values in Table 4 and S5 make it apparent that R^2 is a biased and potentially misleading parameter when used to compare models of differing data. In this application, R^2 is a measure of the proportion of the total variation of peak position from the average peak position in that dataset that is explained by the regression line (McKillip and Dyar 2010). For example, if all the samples in the dataset have comparable %Fo contents, then the deviation from that average will be small, and R^2 may be misleadingly high. Moreover, R^2 cannot be used to evaluate whether the calculated regression function is a correct description of the relationship between peak position and %Fo. Finally, the R^2 statistic does not evaluate potential performance of the linear or polynomial trend when applied to unseen (i.e., from a different dataset or instrument) results.

The importance of this point is reinforced by the data in Table 4, in which only RMSE can be used to make apples-to-apples comparisons among models. For example, Kuebler et al. (2006) modeled %Fo using a second-order polynomial fit to the DB1 and DB2 centroids plotted against %Fo, yielding regression lines with R^2 values of 0.98 (DB1) and 0.97 (DB2) %Fo, respectively. When these data were used to calculate LOO RMSE-CV, a different story emerges. Corresponding LOO RMSE-CV values are ±4.33 and ±4.57 %Fo units. The inaccuracy of this model becomes even more apparent when it is used to predict a different dataset (RMSE-test values). These results underscore the importance of evaluating model accuracies based on use of leave-one-out cross-validation and/or "unseen" external data. Failure to do so invalidates any claims of accuracy for application to other datasets, such as those on Mars.

Future work

In a study using 3,950 RRUFF spectra to test for matching accuracy, Carey et al. (2015) tested the effects of common spectrum pre-processing steps, such as intensity normalization,

smoothing, squashing, and customized baseline removal (Giguere et al. 2017). Although differences in sample crystal orientation, laser polarization, focus, and other instrumental parameters can have major effects on spectra, even on the same samples and identical instruments, Carey et al. (2015) showed that pre-processing techniques can effectively ameliorate these differences and improve mineral identification. It is likely that optimizing pre-processing of olivine spectra from disparate sources might also improve prediction accuracy for obtaining %Fo from Raman spectra of olivine.

However, time did not permit testing of various pre-processing techniques on our own datasets, though this is obviously an area ripe for research. In future work, effects of baseline removal methods, alternate methods for normalization, and squashing and smoothing techniques will be evaluated. Although olivine is an important rock-forming mineral group, there are many other mineral groups that will need to be evaluated and quantitatively described through Raman spectroscopy. Given the ubiquitous presence of species from the pyroxene and feldspar mineral groups in igneous rocks and on planetary surfaces, creation of equivalent multivariate Raman models for these groups should be a high priority for further research. Eventually, bringing our results together with those analogous models for other phases will make it possible to identify and quantify compositions of these phases in mineral mixtures.

IMPLICATIONS

As the technology for micro-Raman, reductions in the price of lasers, and implementations for portable and remote instruments continue, Raman spectroscopy should play an increasingly important role in geosciences and planetary exploration. Studies such as this one that relate prominent Raman peaks to mineral identification and composition are needed to enhance the capabilities of Raman instruments across those many applications. This paper lays a

foundation for future analogous studies of important rock-forming minerals by demonstrating that specific features and energy ranges can be mined for accurate predictions of chemistry. This project focuses on olivine, a common liquidus phase in magmatic systems that is present in most basalts, in meteorites, and on terrestrial surfaces beyond Earth. Characterizing the ratio of Fe to Mg in olivine constrains phase relations and crystallization conditions, and is feasible in both pure spectra of olivine and in mixtures where the most prominent olivine doublet can be resolved. Algorithms presented here will assist the *ExoMars* and *Mars 2020* mission teams in recognizing olivine and determining its composition with known accuracy. We recommend use of either our PLS model if pure olivine is encountered, or the recommended $\%Fo = -0.179625x^2 + 310.077x - 133717$ regression equation (where x = DB2 in units of cm^{-1}) for olivine that is present in mixtures.

ACKNOWLEDGMENTS

We thank the Massachusetts Space Grant Consortium for student funding in support of this project.

REFERENCES CITED

Abdu, Y.A., Varela, M.E., and Hawthorne, F.C. (2011) Raman, FTIR, and Mössbauer spectroscopy of olivines from the D'Orbigny meteorite. Annual Meteoritical Society Meeting, 74, 5112 (abstract).

Allen, F.S., Zhao, J., and Butterfield, D.S. (2000) Apparatus for Measuring and Applying Instrumentation Correction to Produce a Standard Raman Spectrum. 6,141,095.

Banfield, J.M., Dyar, M.D., and McGuire, A.V. (1992) The defect microstructure of oxidized mantle olivine from Dish Hill, California. American Mineralogist, 77, 959-975.

Breitenfeld, L.B. (2017) Predicting olivine composition using Raman spectroscopy through band shift and multivariate analyses, 55-56 p. B.A. thesis, Mount Holyoke College, South Hadley.

Butler, N.A., and Denham, M.C. (2000) The peculiar shrinkage properties of partial least squares regression. Journal of the Royal Statistical Society, 62, 585-593.

Byrne, S.A., Dyar, M.D., Bessette, E.E., Breitenfeld, L.B., Crowley, M.C., Hoff, C.M., Marchand, G.J., Ketley, M.N., Roberts, A.L., Sklute, E.C., and Parente, M. (2015) Pure mineral separates for mixing experiments to simulate planetary surfaces. Lunar and Planetary Science Conference, 46, 1499 (abstract).

Carey, C., Dyar, M.D., Boucher, T., and Mahadevan, S. (2015) Machine learning tools for mineral recognition and classification from Raman spectroscopy. Journal of Raman Spectroscopy, 46, 894-903.

Carey, C., Dyar, M.D., Boucher, T., and Giguere, S. (2017) Web-based software for preprocessing, matching, fitting, prediction, and visualization of spectroscopic data: the data exploration, visualization, and analysis of spectra (DEVAS) website. Lunar and Planetary Science Conference, 48, 1097 (abstract).

Chopelas, A. (1991) Single crystal Raman spectra of forsterite, fayalite, and monticellite. American Mineralogist, 76, 1101-1109.

Dyar, M.D. (2003) Ferric iron in SNC meteorites as determined by Mössbauer spectroscopy: Implications for Martian landers and Martian oxygen fugacity. Meteoritics & Planetary Science, 38, 1733-1752.

Dyar, M.D., McGuire, A.V., and Ziegler, R.D. (1989) Redox equilibria and crystal chemistry of coexisting minerals from spinel lherzolite mantle xenoliths. American Mineralogist, 74,

969-980.

Dyar, M.D., McGuire, A.V., and Harrell, M.D. (1992) Crystal chemistry of iron in two styles of metasomatism in the upper mantle. Geochimica et cosmochimica acta, 56, 2579-2586.

Dyar, M.D., Delaney, J.S., Sutton, S.R., and Schaefer, M.W. (1998) Fe^{3+} distribution in oxidized olivine: A synchrotron micro-XANES study. American Mineralogist, 83, 1361-1365.

Dyar, M.D., Carmosino, M.L., Speicher, EA., Ozanne, M.V., Clegg, S.M., and Wiens, R.C. (2012a) Comparison of partial least squares and lasso regression techniques for laser-induced breakdown spectroscopy of geological samples. Spectrochimica Acta B, 70, 51-67.

Dyar, M.D., Fassett, C.I., Giguere, S., Lepore, K., Byrne, S., Boucher, T., Carey, C., and Mahadevan, S. (2016a) Comparison of univariate and multivariate models for prediction of major and minor elements from laser-induced breakdown spectra with and without making. Spectrochimica Acta B, 123, 93-104.

Dyar, M.D., Breves, E.A., Gunter, M.E., Lanzirotti, A., Tucker, J.M., Carey, C., Peel, S.E., Brown, E.B., Oberti, R., Lerotic, M., and Delaney, J.S. (2016b) Use of multivariate analysis for synchrotron micro-XANES analysis of iron valence state in amphiboles. American Mineralogist, 101, 1171-1189.

Dyar, M.D., Breitenfeld, L.B., Carey, CJ, Bartholomew, P., Tague, T.J., Wang,P., Mertzman, S., Byrne, S.A., Crowley, M.C., Leight, C., Watts, E., Campbell, J.C., Celestian, A., McKeeby, B., Jaret, S., Glotch, T., Berlanga, G., and Misra, A. (2016c) Interlaboratory and cross-instrument comparison of Raman spectra of 96 minerals. Lunar and Planetary Science 47, Abstract #2240.

Dyar, M.D., Breves, E.A., Emerson, E., Bell, S.W., Nelms, M., Ozannes, M.V., Peel, S.E.,

564 Carmosino, M.L., Tucker, J.M., Gunter, M.E., Delaney, J.S., Lanzirotti, A., and
565 Woodland, A.B. (2012b) Accurate determination of ferric iron in garnets by bulk
566 Mössbauer spectroscopy and synchrotron micro-XANES. American Mineralogist, 97,
567 1726-1740.
568 Dyar, M.D., Sklute, E.C., Menzies, O.N., Bland, P.A., Lindsley, D., Glotch, T., Lane, M.D.,
569 Wopenka, B., Klima, R., Bishop, J.L., Hiroi, T., Pieters, C.M., and Sunshine, J. (2009)
570 Spectroscopic characteristics of synthetic olivines, with an emphasis on fayalite: An
571 integrated multi-wavelength approach. American Mineralogist, 94, 883-898.
572 Floran, R.J., Prinz, M., Hlava, P.F., Keil, K., Nehru, C.E., and Hinthorne, J.R. (1978) The
573 Chassigny meteorite: A cumulate dunite with hydrous amphibole-bearing melt inclusions.
574 Geochimica et Cosmochimica Acta, 42, 1213-1229.
575 Foster, N.J., Wozniakiewicz, P.J., Burchell, M.J., Kearsley, A.T., Creighton, A.J., and Cole, M.J.
576 (2007) Identification by Raman spectroscopy of processing effects in forsterite- fayalite
577 samples during hypervelocity impacts on foils and capture in aerogel. Annual
578 Meteoritical Society Meeting, 70, 5186 (abstract).
579 Gaisler, S.V., and Kolesov, B.A. (2007) Raman spectra of olivine solid solution $(Fe_xMg_{1-x})_2SiO_4$
580 and spin-vibration interaction. Journal of Structural Chemistry, 48, 61-65.
581 Giguere, S., Boucher, T., Carey, CJ, Mahadevan, S., and Dyar, M.D. (2017) A fully-customized
582 baseline removal framework for spectroscopic applications. Applied Spectroscopy, 71,
583 1457-1470.
584 Guyot, F., Boyer, H., Madon, M., Velde, B., and Poirier, J.P. (1986) Comparison of the Raman
585 microprobe spectra of $(Mg, Fe)_2SiO_4$ and Mg_2GeO_4 with olivine and spinel structures.
586 Physics and Chemistry of Minerals, 13, 91-95.

587 Hastie, T., Tibshirani, R., and Friedman, J. (2009) The Elements of Statistical Learning, 745 p.
588 Springer Science, New York.
589 Iishi, K. (1978) Lattice dynamics of forsterite. American Mineralogist, 63, 1198-1208.
590 Isaacson, P.J., Klima, R.L., Sunshine, J.M., Pieters, C.M., Hiroi, T., and Dyar, M.D. (2014)
591 Visible to near-infrared reflectance spectroscopy of pure synthetic olivine across the
592 olivine solid solution. American Mineralogist, 99, 467-478.
593 Ishibashi, H., Arakawa, M., Yamamoto, J., and Kagi, H. (2011) Precise determination of Mg/Fe
594 ratios applicable to terrestrial olivine samples using Raman spectroscopy. Journal of
595 Raman Spectroscopy, 43, 331-337.
596 Ishibashi, H., Arakawa, M., Ohi, S., Yamamoto, J., Miyake, A., and Kagi, H. (2008)
597 Relationship between Raman spectral pattern and crystallographic orientation of a rock-
598 forming mineral: a case study of $Fo_{89}Fa_{11}$ olivine. Journal of Raman Spectroscopy, 39,
599 1653-1659.
600 Kolesov, B.A., and Geiger, C.A., (2004) A Raman spectroscopic study of Fe–Mg olivines.
601 Physics and Chemistry of Minerals, 31, 142–154.
602 Kolesov, B.A., and Tanskaya, J.V. (1996) Raman spectra and cation distribution in the lattice of
603 olivines. Materials Research Bulletin, 31, 1035-1044.
604 Kuebler, K.E., Jolliff, B.L., Wang, A., and Haskin, L.A. (2006) Extracting olivine (Fo-Fa)
605 compositions from Raman spectral peak positions. Geochimica et Cosmochimica Acta,
606 70, 6201-6222.
607 Lane, M.D., Glotch, T.D., Dyar, M. D., Pieters, C.M., Klima, R., Hiroi, T., Bishop, J.L., and
608 Sunshine, J. (2011) Midinfrared spectroscopy of synthetic olivines: Thermal emission,
609 specular and diffuse reflectance, and attenuated total reflectance studies of forsterite to

610 fayalite. Journal of Geophysical Research, 116, E08010, DOI: 10.1029/2010JE003588.
611 McCanta, M.C., Treiman, A.H., Dyar, M.D., Alexander, C.M.O'D., Rumble, D., and Essene,
612 E.J. (2008) The LaPaz Icefield 04840 meteorite: Mineralogy, metamorphism, and origin
613 of an amphibole-and biotite-bearing R chondrite. Geochimica et Cosmochimica Acta, 72,
614 5757-5780.
615 McGuire, A.V., Dyar, M.D., and Nielson, J.E. (1991) Metasomatic oxidation of upper mantle
616 periodotite. Contributions to Mineralogy and Petrology, 109, 252-264.
617 McGuire, A.V., Francis, C.A., and Dyar, M.D. (1992). Mineral standards for electron
618 microprobe analysis of oxygen. American Mineralogist, 77, 1087-1087.
619 McKeown, D.A., Bell, M.I., and Caracas, R. (2010) Theoretical determination of the Raman
620 spectra of single-crystal forsterite (Mg_2SiO_4). American Mineralogist, 95, 980-986.
621 McKillup, S., and Dyar, M.D. (2010) Geostatistics explained: An introductory guide for earth
622 scientists, 396 p. Cambridge University Press, Cambridge, UK.
623 McSween, H.Y., and Jarosewich, E. (1983) Petrogenesis of the Elephant Moraine A79001
624 meteorite: Multiple magma pulses on the shergottite parent body. Geochimica et
625 Cosmochimica Acta, 47, 1501-1513.
626 Mikouchi, T., and Kurihara, T. (2008) Mineralogy and petrology of paired lherzolitic shergottites
627 Yamato 000027, Yamato 000047, and Yamato 000097: Another fragment from a Martian
628 "lherzolite" block. Polar Science, 2, 175-194.
629 Mohanan, K., Sharma, S.K., and Bishop F.C. (1993) A Raman spectral study of forsterite-
630 monticellite solid solutions. American Mineralogist, 78, 42-48.
631 Moral, A.G., Rull, F., Maurice, S., Hutchinson, I., Canora, C.P., Seoane, L., Canchal, R.,
632 Gallego, P., Ramos, G., Prieto, J.A.R., Santiago, A., Santamaría, P., Colombo, M.,

633 Belenguer, T., López, G., Quintana, C., Zafra, J., Berrocal, A., Pintor, C., Cabrero, J., and
634 Saiz, J, (2018) Raman laser spectrometer for 2020 ExoMars Mission. Engineering and
635 qualification model capabilities and future activities. Lunar and Planetary Science
636 Conference, 49, 2449 (abstract).
637 Morse, S.A. (1996) Kiglapait mineralogy .3. Olivine compositions and Rayleigh fractionation
638 models. Journal of Petrology, 37, 1037-1061.
639 Morse, S.A. (2001) Augite-olivine equilibria in the Kiglapait intrusion, Labrador, Canada.
640 Canadian Mineralogist, 39, 267-274.
641 Mouri, T., and Enami, M. (2008) Raman spectroscopic study of olivine-group minerals. Journal
642 of Mineralogical and Petrological Sciences, 103, 100-104.
643 Mullen, T., Dyar, M.D., Parente, M., and Breitenfeld, L.B. (2018) Improving matching accuracy
644 in Raman spectroscopy by quantifying the wavenumber shift between instruments. Lunar
645 and Planetary Science, 49, Abstract #1185.
646 Parques-Ledent, M.T., and Tarte, P. (1973) Vibrational studies of olivine-type compound-I. The
647 i.r. and Raman spectra of the isotopic species of Mg_2SiO_4. Spectrochimica Acta A, 29,
648 1007-1016.
649 Piriou, B. (1983) The high-frequency vibrational spectra of vitreous and crystalline
650 orthosilicates. American Mineralogist, 68, 426-443.
651 Price, G.D., Parker, S.C., and Leslie, M., (1987) The lattice dynamics of forsterite. Mineralogical
652 Magazine 51, 157-170.
653 Schaefer, M.W. (1983) Crystal chemistry of ferric-rich fayalites, 38-56 p. Ph.D. thesis, MIT,
654 Cambridge.
655 Servoin, J., Piriou, B., and Alain, P. (1972) Diffusion Raman de la forsterite sythetique pure,

Mg_2SiO_4. Comptes Rendus de l'Académie des Sciences Paris, 274, 135-137.

Sklute, E.C. (2006) Mössbauer spectroscopy of synthetic olivine across the Fe-Mg solid solution, 1-38 p. B.A. thesis, Mount Holyoke College, South Hadley.

Stone, M., and Brooks, R.J. (1990) Continuum regression: cross-validated sequentially constructed predictions embracing ordinary least-squares, partial least-squares, and principal components regression. Journal of the Royal Statistical Society B, 52, 237–269.

Tibshirani, R. (1995) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58, 267-288.

Tucker, J.M., Dyar, M.D., Schaefer, M.W., Clegg, S.M., and Wiens, R.C. (2010) Optimization of laser-induced breakdown spectroscopy for rapid geochemical analysis. Chemical Geology, 277, 137-148.

Wang, A., Kuebler, K., Jolliff, B., and Haskin, L.A. (2004) Mineralogy of a Martian meteorite as determined by Raman spectroscopy. Journal of Raman Spectroscopy, 35, 504-514.

Wiens, R.C., Newell, R., Clegg, S., Sharma, S.K., Misra, A., Bernardi, P., Maurice, S., McCabe, K., Cais, P., and the SuperCam Science Team (2017) The SuperCam remote Raman spectrometer for Mars 2020. Lunar and Planetary Science, 48, 2600 (abstract).

Wold, S., Martens, H., and Wold, H. (1983) The multivariate calibration problem in chemistry solved by the PLS method. Lecture Notes in Mathematics, 973, 286-293.

Yasuzuka, T., Ishibashi, H., Arakawa, M., Yamamoto, J., and Kagi, H. (2009) Simultaneous determination of Mg# and residual pressure in olivine using micro-Raman spectroscopy. Journal of Mineralogical and Petrological Sciences, 104, 395-400.

Zhang, Z.M., Chen, S., and Liang, Y.Z. (2010) Baseline correction using adaptive iteratively reweighted penalized least squares. Royal Society of Chemistry, 135, 1138-1146.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

679

680 Figure captions

681 Figure 1. High intensity Raman doublet (815-857 cm^{-1}) of forsterite (purple) and fayalite (green).

682 Figure 2. Unaligned Raman spectra of olivine doublet (DB1 and DB2) of 93 samples acquired on
683 Bruker's Senterra and BRAVO spectrometers. All spectra were baseline removed using Air-PLS
684 and normalized to a maximum intensity of 1. Spectra are color-coded based on Fo content, where
685 forsterite is represented with yellow, fayalite with purple, and intermediate compositions in
686 between.

687 Figure 3. Histogram of 93 synthetic (blue) and natural (red) samples on the Fo-Fa series. Natural
688 olivines typically form with a %Fo of ~89.5 resulting in an unbalanced distribution on the Fo-Fa
689 series.

690 Figure 4. Fo by EMPA versus peak centroids of DB1 and DB2. Second order polynomial fits
691 and RSME-CV values are included for the unaligned data acquired on Bruker's BRAVO (*n*=25)
692 and Senterra (*n*=68) spectrometers. Error bars are smaller than the symbols and are given in
693 Tables S3 and S4.

694 Figure 5. %Fo by EMPA versus peak centroids positions of (a) DB1 and (b) DB2 for data
695 acquired in other studies. Including results from the RRUFF* database (see text for explanation
696 of notation), and studies by Kuebler et al. (2006), Yasuzuka et al. (2009), and Ishibashi et al.
697 (2011).

698 Figure 6. Variations in Lasso model accuracy as a function of the number of coefficients. As α
699 increases, fewer channels are examined: (a) 38 channels for $\alpha = 0.001$, (b) 27 channels for $\alpha =$
700 0.01, and (c) 10 channels for $\alpha = 0.1$. As the number of channels examined is decreased (fewer

coefficients within the model), the RMSE-CV of the models increases in value (gets worse). This demonstrates the value of a models that examines a high number of channels, which is achieved in a small α value Lasso model or PLS models.

Figure 7. Bar graph comparing results from Tables 4 and 5. When the dataset is small and/or all the data are acquired on the identical instrument, then univariate methods produce better results than those using multivariate analyses. However, as the number of samples and instruments used increase, PLS methods generally produce more accurate results.

Figure 8. (top) Plot of BRAVO and Senterra unaligned data in Table 6, along with circles indicating the magnitude of PLS coefficients (right-hand y axis units). Note that PLS coefficients are proportional to spectral intensity at each wavenumbers, so absolute values cannot be compared on this plot. However, the PLS coefficients do demonstrate that the entire wavenumber range contains information useful to predicting composition. (bottom) Analogous plot for the same data but showing Lasso coefficients for a model with $\alpha = 0.001$.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Table 1. Assignment of Raman Modes in Olivine

~Band (cm^{-1})	Symmetry	Assignment	Literature
815-825 (DB1)	A_g	n.a.	Servoin et al. (1972)
		v_1+v_3	Paques-Ledent and Tarte (1973)
		v_1	Iishi (1978)
		$v_1 +v_3$	Piriou (1983)
		v_1+v_3	Chopelas et al. (1991)
		SiO_4^{2-} stretching	Kolesov and Tanskaya (1996)
		v_1+v_3	Kolesov and Geiger (2004)
		Si-O stretch, v_3	McKeown et al. (2010)
838	B_{1g}	v_1	Iishi (1978)
		v_1 (+v_3)	Chopelas et al. (1991)
		v_1 (+v_3)	Kolesov and Geiger (2004)
		v_1	McKeown et al. (2010)
837-857 (DB2)	A_g	n.a.	Servoin et al. (1972)
		v_1+v_3	Paques-Ledent and Tarte (1973)
		v_3	Iishi (1978)
		v_1 (+v_3)	Piriou (1983)
		v_3	Price et al. (1987)
		v_1+v_3	Chopelas et al. (1991)
		SiO_4^{2-} stretching	Kolesov and Tanskaya (1996)
		v_1+v_3	Kolesov and Geiger (2004)
		Si-O stretch; SiO_4 breathing v_3	McKeown et al. (2010)
866	B_{1g}	v_3	Iishi (1978)
		v_3	Price et al. (1987)
		v_3 (+v_1)	Chopelas et al. (1991)
		v_3 (+v_1)	Kolesov and Geiger (2004)
882	B_{2g}	v_3	McKeown et al. (2010)
		v_3	Paques-Ledent and Tarte (1973)
		v_3	Iishi (1978)
		v_3	Chopelas et al. (1991)
		v_3	Kolesov and Geiger (2004)
		v_3	McKeown et al. (2010)

1

2

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Table 2. Summary of %Fo prediction models using band shift method

Paper	**# samples***	**Bands (cm^{-1})**
Iishi (1978)	1	All bands
Guyot et al. (1986)	4	815-825, 837-857, 881-883, 914-920, 950-966
Chopelas et al. (1991)	1	All bands
Mohanan et al. (1993)	1	All bands from 200-1000
Kolesov and Tanskaya, (1996)	2	All bands from 200-1000
Wang et al. (2004)	0	815-825, 837-857
Kuebler et al. (2006)	10	815-825, 837-857
Foster et al. (2007)	2	815-825, 837-857
Gaisler and Kolesov (2007)	0	200-230, 290-310, 410-440, 815-825, 837-857
Mouri and Enami (2008)	0	815-825, 837-857
Ishibashi et al. (2008)	1	815-825, 837-857
Yasuzuka et al. (2009)	10	540-553, 815-825, 837-857
McKeown et al. (2010)	1	All bands
Abdu et al. (2011)	3	815-825, 837-857
Ishibashi et al. (2011)	15	815-825, 837-857, 881-883, 914-920, 950-966
Weber et al. (2014)	5	815-825, 837-857

Note: *Number of samples with reported centroid positions and compositions.

3

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Table 3. Natural Samples Studied

Sample Name	Locality	Chemistry	Mössbauer	Raman Instrument
Ba-1-61	Dish Hill, CA	UTK	[12]	Senterra
Ba-1-74	Dish Hill, CA	UTK	[12]	Senterra
Ba-2-1 WR1	Dish Hill, CA	Brown	[6]	Senterra
Ba-2-1 WR2	Dish Hill, CA	Brown	[6]	Senterra
Ba-2-1 WR3	Dish Hill, CA	Brown	[6]	Senterra
Ba-2-1 WR4	Dish Hill, CA	Brown	[6]	Senterra
Ba-2-1 D-1	Dish Hill, CA	[1]	[6]	Senterra
Ci-1-183	Dish Hill, CA	[12]	[12]	Senterra
Ci-1-25	Dish Hill, CA	[12]	[12]	Senterra
DH101-B	Dish Hill, CA	Brown	[6]	Senterra
DH101-C	Dish Hill, CA	Brown	[6]	Senterra
DH101-D	Dish Hill, CA	Brown	[6]	Senterra
DH101-E	Dish Hill, CA	Brown	[6]	Senterra
Dyar 89-190	unknown	Brown	[12]	BRAVO
Dyar 89-12	unknown	Brown	[12]	BRAVO
Dyar 89-187	unknown	Brown	[12]	BRAVO
Dyar 89-194	unknown	Brown	[12]	BRAVO
Ep-1-13	Potrillo maar, NM	[12]	[7]	BRAVO
Ep-3-139-C	Kilbourne Hole, NM	Brown	[8]	Senterra
Ep-3-139-D	Kilbourne Hole, NM	Brown	[8]	Senterra
Ep-3-44	Kilbourne Hole, NM	UTK	[12]	Senterra
Ep-3-46	Kilbourne Hole, NM	UTK	[12]	Senterra
Ep-3-72	Kilbourne Hole, NM	UTK	[12]	Senterra
Ep-3-7A	Kilbourne Hole, NM	Univ. Houston	[12]	Senterra
KI-3003	Kiglapait Formation	Brown	[12]	Senterra
KI-3373	Kiglapait Formation	Brown	[12]	Senterra
NMNH 112085	Red Rock Ridge	UTK, Brown	[12]	Senterra
NMNH 1210672	Germany Greifensteiner Kalk	UTK, Brown	[12]	Senterra
NMNH 135841	Sweden Nykopig Tunaberg	Brown	[12]	Senterra
NMNH 85539	unknown	UTK, Brown	[12]	Senterra
Rockport	Rockport	Brown	[9]	Senterra
Globe	Globe, AZ	[1]	[12]	BRAVO
H279-12	Harrat al Kishb, Saudi Arabia	[12]	[12]	Senterra
H30-82-8	Harrat al Kishb, Saudi Arabia	UTK	[12]	Senterra
H30-B1	Harrat al Kishb, Saudi Arabia	Brown	[10]	BRAVO
H30-B2	Harrat al Kishb, Saudi Arabia	[12]	[7]	BRAVO
H30-B3	Harrat al Kishb, Saudi Arabia	UTK	[12]	Senterra
H30-B4	Harrat al Kishb, Saudi Arabia	UTK, Brown	[12]	Senterra
H30-B5	Harrat al Kishb, Saudi Arabia	UTK	[12]	Senterra
H312-1	Harrat Uwayrid, Saudi Arabia	[12]	[12]	Senterra
H366-28	Harrat Hutaymah, Saudi Arabia	[12]	[12]	BRAVO
H366-30	Harrat Hutaymah, Saudi Arabia	[12]	[12]	Senterra
NMNH 9140	Orange Co. NY	UTK, Brown	[12]	Senterra
KBH-94-23-B	Kilbourne Hole, NM	UTK	[12]	Senterra
KBH-94-23-E	Kilbourne Hole, NM	UTK	[12]	Senterra

Notes: Sources are abbreviated as follows: [1] Byrne et al. (2015), [2] Floran et al. (1978), [3] McSween and Jarosewich (1983), [4] McCanta et al. (2008), [5] Mikouchi and Kurihara (2008), [6] McGuire et al. (1991), [7] Dyar et al. (1989), [8] Dyar et al.(1992), [9] Schaefer (1983), [10] McGuire et al. (1992), [11] Dyar (2003), [12] this study.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Table 3. (continued). **Natural Samples Studied**

Sample Name	Locality	Chemistry	Mössbauer	Raman Instrument
KBH-94-23-E	Kilbourne Hole, NM	UTK	[12]	Senterra
KI-3005	Kiglapait Formation	Brown	[12]	Senterra
KI-3289	Kiglapait Formation	Brown	[12]	Senterra
KI-3362	Kiglapait Formation	Brown	[12]	Senterra
KI-3648	Kiglapait Formation	UTK	[12]	Senterra
KI-4030	Kiglapait Formation	Brown	[12]	Senterra
Ki-5-16	Cima volcanic field, CA	[12]	[12]	Senterra
Ki-5-235	Cima volcanic field, CA	UTK	[12]	Senterra
Ki-5-35	Cima volcanic field, CA	UTK	[12]	Senterra
Ki-5-62	Cima volcanic field, CA	UTK	[12]	BRAVO
Pakistan	Pakistan Sapatime Kohistan District	Brown	[12]	BRAVO
San Carlos AZ	San Carlos AZ	[1]	[12]	BRAVO
ALHA 77005	Mars	UTK	[11]	Senterra
ALHA-77005-193	Mars	UTK	[11]	Senterra
Chassigny USNM E24	Mars	[2]	[11]	Senterra
EETA-79001 60B	Mars	[3]	[11]	Senterra
EETA-79001-A	Mars	[3]	[11]	Senterra
LAP-0484016	Mars	[4]	[12]	Senterra
NWA2737	Mars	Brown	[12]	Senterra
Y000097 86	Mars	[5]	[12]	Senterra

Notes: Sources are abbreviated as follows: [11] Dyar (2003), [12] this study.

6

7

Table 4. Model accuracy by LOO RMSE-CV for univariate analyses using 2nd order polynomial fits

Data	#Spectra	Resolution (cm^{-1} /channel)	Model	Internal R^2	Internal RMSE	LOO RMSE-CV
BRAVO	25	2.0	DB1	0.88	9.87	7.69
			DB2	0.97	5.26	4.35
Senterra	68	0.5	DB1	0.94	7.38	4.94
			DB2	0.97	5.20	3.35
Senterra + BRAVO	93	0.5 / 2.0	DB1	0.89	10.23	7.43
			DB2	0.89	10.20	7.91
Senterra + BRAVO aligned	93	0.5 / 2.0	DB1	0.92	8.72	5.64
			DB2	0.94	7.36	5.05
Kuebler et al. (2006)	13	6.2	DB1	0.98	8.63	4.33
			DB2	0.97	5.05	4.57
Yasuzuka et al. (2009)	10	0.05	DB1	0.97	1.49	1.68
			DB2	0.97	1.57	1.84
Ishibashi et al. (2011)	15	1.5	DB1	0.98	1.30	1.32
			DB2	0.99	0.92	0.84
RRUFF*	156	0.48 - 1.4	DB1	0.92	11.58	8.75
			DB2	0.94	9.93	8.28
RRUFF (all)	188	2.0	DB1	0.86	15.50	11.29
			DB2	0.93	10.70	8.86
Senterra + BRAVO+RRUFF*	249	0.5/2.0 /0.48 – 1.4	DB1	0.90	12.04	9.36
			DB2	0.92	10.58	8.91
Senterra + BRAVO + RRUFF (all)	281	0.5/2.0 /0.48 – 2.0	DB1	0.86	14.35	10.55
			DB2	0.92	11.02	9.20
Senterra + BRAVO+RRUFF* (>50%Fo)	181	0.5/2.0 /0.48 – 1.4	DB1	0.56	6.60	5.01
			DB2	0.61	6.20	4.58
All data (only RRUFF*)	287	see above	DB1	0.90	11.49	8.70
			DB2	0.92	10.26	8.56
All data for samples >50%Fo (only RRUFF*)	213	see above	DB1	0.64	6.30	4.86
			DB2	0.69	5.85	4.27

*Includes all olivine group spectra on RRUFF except those designated as "Broad Scan with Spectral Artifacts."

8

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

Table 5. Model accuracy by LOO RMSE-CV for Multivariate Analyses

Data	#	Res.	Model	σ	C/α	Internal R^2	Internal RMSE	LOO RMSE-CV
BRAVO	25	3.0	PLS	800-880	7	0.99	2.56	**6.85**
		3.0	Lasso	800-880	0.001	0.99	1.13	6.87
		2.0	PLS	800-880	2	0.96	5.75	7.43
		2.0	Lasso	800-880	0.005	0.99	1.88	9.24
		2.0	PLS	5 regions	9	0.99	1.67	8.17
		2.0	Lasso	5 regions	0.3	0.93	7.13	7.91
Senterra	68	3.0	PLS	800-880	6	0.97	5.57	7.06
		3.0	Lasso	800-880	0.01	0.93	8.52	10.62
		0.5	PLS	800-880	7	0.96	5.85	7.31
		0.5	Lasso	800-880	0.008	0.95	7.37	11.75
		2.0	PLS	5 regions	9	0.98	4.40	**6.45**
		2.0	Lasso	5 regions	0.005	0.98	3.93	7.50
Senterra + BRAVO	93	2.0	PLS	800-880	5	0.95	6.63	**7.79**
		2.0	Lasso	800-880	0.002	0.95	6.95	9.38
		2.0	PLS	5 regions	9	0.96	5.79	7.91
		2.0	Lasso	5 regions	0.006	0.97	5.06	8.72
Senterra + BRAVO aligned	93	2.0	PLS	800-880	7	0.97	5.69	**6.93**
		2.0	Lasso	800-880	0.003	0.95	7.19	9.42
		2.0	PLS	5 regions	8	0.96	5.87	7.90
		2.0	Lasso	5 regions	0.02	0.95	7.07	9.74
	118	2.0	PLS	800-880	7	0.97	5.56	**6.63**
		2.0	Lasso	800-880	0.002	0.94	7.34	9.52
		2.0	PLS	5 regions	9	0.96	5.78	7.59
		2.0	Lasso	5 regions	0.015	0.95	7.03	8.96
RRUFF*	156	1.5	PLS	800-880	5	0.98	6.10	6.65
		1.5	Lasso	800-880	0.01	0.94	9.97	12.25
		1.5	PLS	5 regions	9	0.99	4.56	**5.97**
		1.5	Lasso	5 regions	0.008	0.98	6.46	7.02
RRUFF (all)	188	2.0	PLS	800-880	10	0.97	7.13	**8.34**
		2.0	Lasso	800-880	0.001	0.93	11.00	13.15
		2.0	PLS	5 regions	10	0.97	7.30	8.53
		2.0	Lasso	5 regions	0.001	0.96	7.85	10.99
Senterra + BRAVO + RRUFF*	249	2.0	PLS	800-880	10	0.97	6.97	**7.83**
		2.0	Lasso	800-880	0.002	0.97	6.97	11.94
		2.0	PLS	5 regions	9	0.96	7.12	8.34
		2.0	Lasso	5 regions	0.002	0.97	7.06	11.29
Senterra + BRAVO + RRUFF (all)	281	2.0	PLS	800-880	9	0.95	8.87	9.64
		2.0	Lasso	800-880	0.001	0.91	11.77	13.41
		2.0	PLS	5 regions	10	0.95	8.57	**9.45**
		2.0	Lasso	5 regions	0.015	0.93	10.08	12.14
Senterra + BRAVO + RRUFF* (>50%Fo)	181	2.0	PLS	800-880	6	0.82	4.23	4.86
		2.0	Lasso	800-880	0.004	-0.28	11.25	13.74
		2.0	PLS	5 regions	9	0.89	3.30	**4.22**
		2.0	Lasso	5 regions	0.02	0.24	8.65	13.28

Notes: # = number of spectra, Res. = Resolution (cm^{-1} /channel), σ = wavenumber range (cm^{-1}), C/a = number of components for PLS models and a for Lasso models. Five regions = 410-440 cm^{-1}, 538-556 cm^{-1}, 800-880 cm^{-1}, 908-926 $cm^{-1}$and 950-968 cm^{-1} combined. *Includes all olivine group spectra on RRUFF except those designated as "Broad Scan with Spectral Artifacts."

9

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

10

Table 6. Comparison of Internal RMSE Results by Energy Region with 2.0 cm^{-1} /channel resolution

	BRAVO		**Senterra**		**BRAVO + Senterra**	
Number of spectra	**25**		**68**		**93**	
Model used	**PLS**	**Lasso**	**PLS**	**Lasso**	**PLS**	**Lasso**
410-440 cm^{-1}	18.96	12.70	15.30	16.46	20.87	20.54
538-556 cm^{-1}	20.82	19.43	22.60	34.09	23.11	31.77
800-880 cm^{-1}	5.75	1.88	5.44	6.04	6.63	6.95
908-926 cm^{-1}	16.04	14.46	12.91	14.74	14.45	16.14
950-968 cm^{-1}	18.24	15.38	19.11	20.97	19.99	24.61
5 regions above	1.67	7.13	4.40	3.93	5.79	5.06
400-700 cm^{-1}	17.34	16.99	8.01	6.37	13.07	12.74
300-1500 cm^{-1}	3.87	6.76	4.52	9.42	5.48	3.74

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

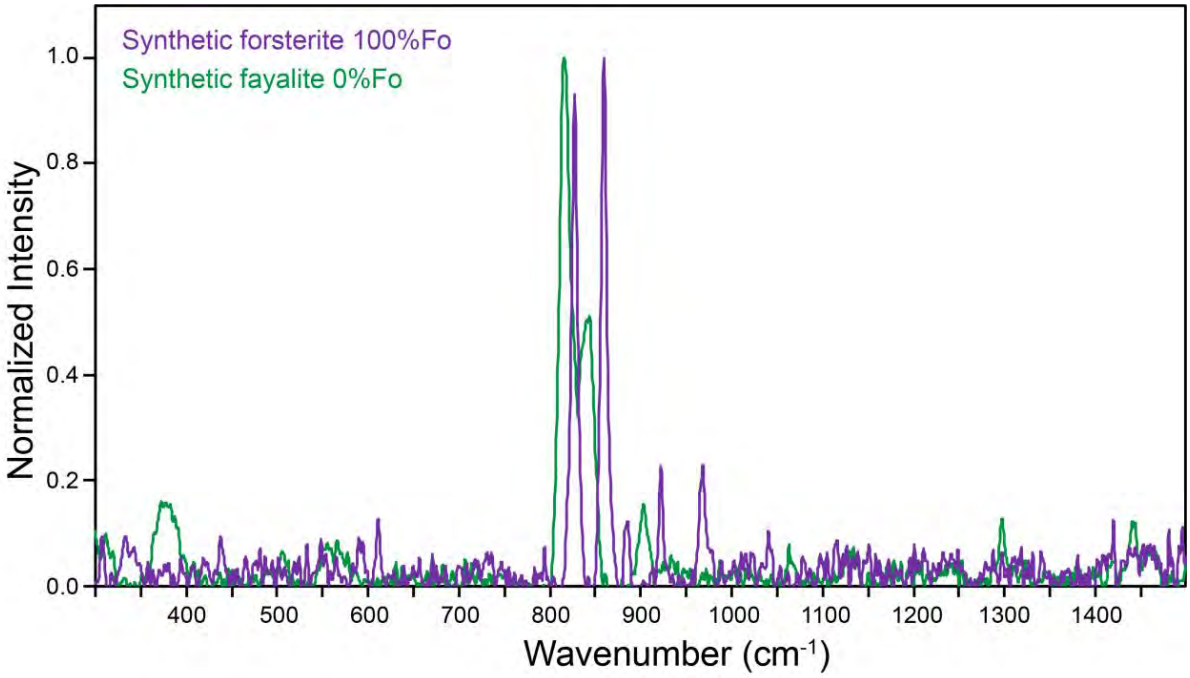


Figure 1. High intensity Raman doublet (815-857 cm^{-1}) of forsterite (purple) and fayalite (green).

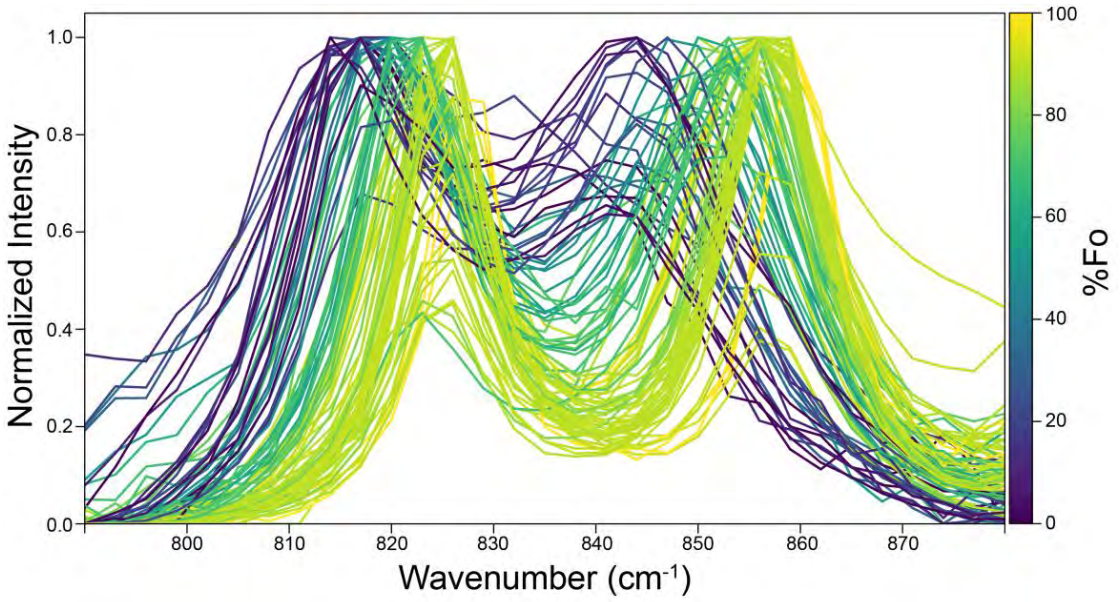


Figure 2. Unaligned Raman spectra of olivine doublet (DB1 and DB2) of 93 samples acquired on Bruker's Senterra and BRAVO spectrometers. All spectra were baseline removed using Air-PLS and normalized to a maximum intensity of 1. Spectra are color-coded based on Fo content, where forsterite is represented with yellow, fayalite with purple, and intermediate compositions in between.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

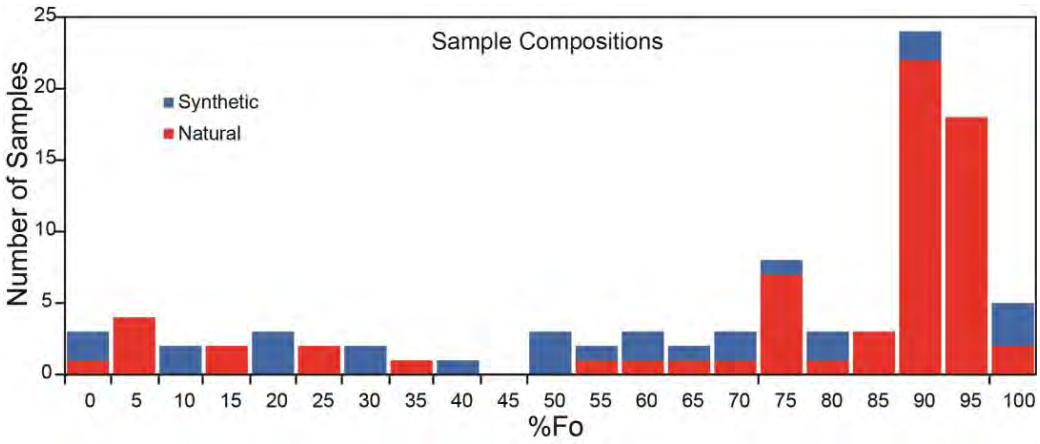
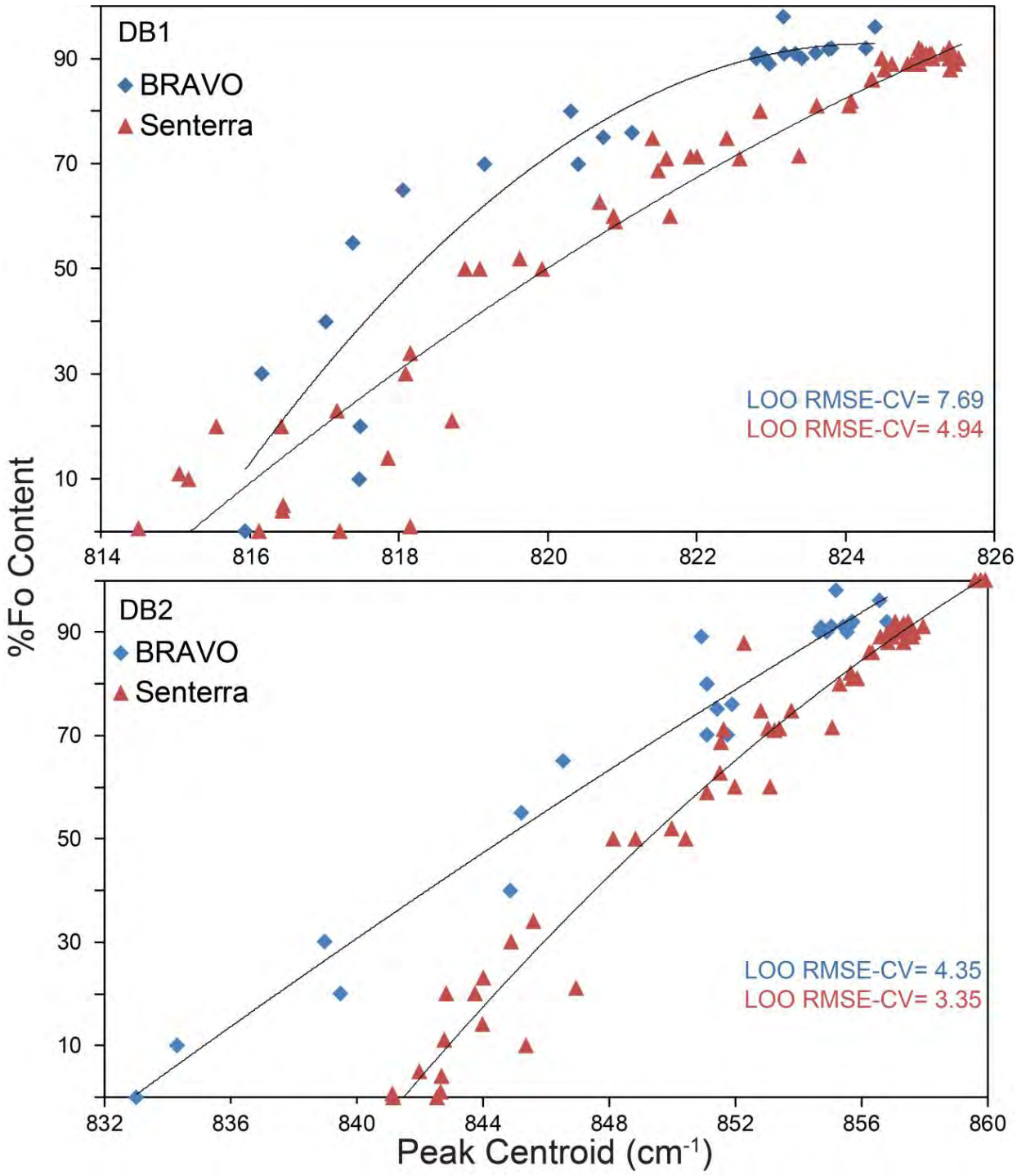


Figure 3. Histogram of 93 synthetic (blue) and natural (red) samples on the Fo-Fa series. Natural olivines typically form with a %Fo of ~89.5 resulting in an unbalanced distribution on the Fo-Fa series.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

23



24

25 Figure 4. Fo by EMPA versus peak centroids of DB1 and DB2. Second order polynomial fits
26 and RSME-CV values are included for the unaligned data acquired on Bruker's BRAVO (*n*=25)
27 and Senterra (*n*=68) spectrometers. Error bars are smaller than the symbols and are given in
28 Tables S3 and S4.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

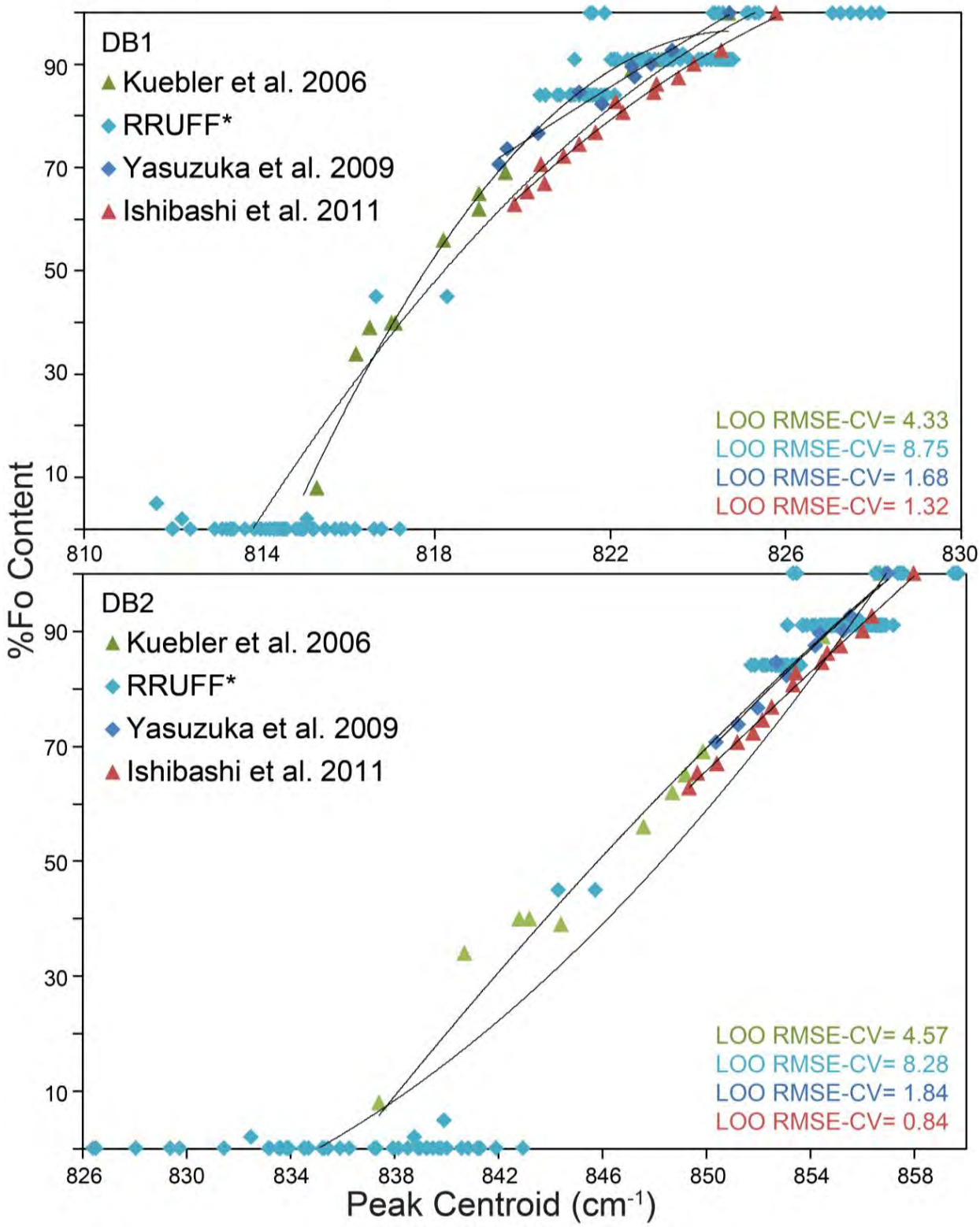


Figure 5. %Fo by EMPA versus peak centroids positions of (a) DB1 and (b) DB2 for data acquired in other studies. Including results from the RRUFF* database (see text for explanation of notation), and studies by Kuebler et al. (2006), Yasuzuka et al. (2009), and Ishibashi et al. (2011).

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

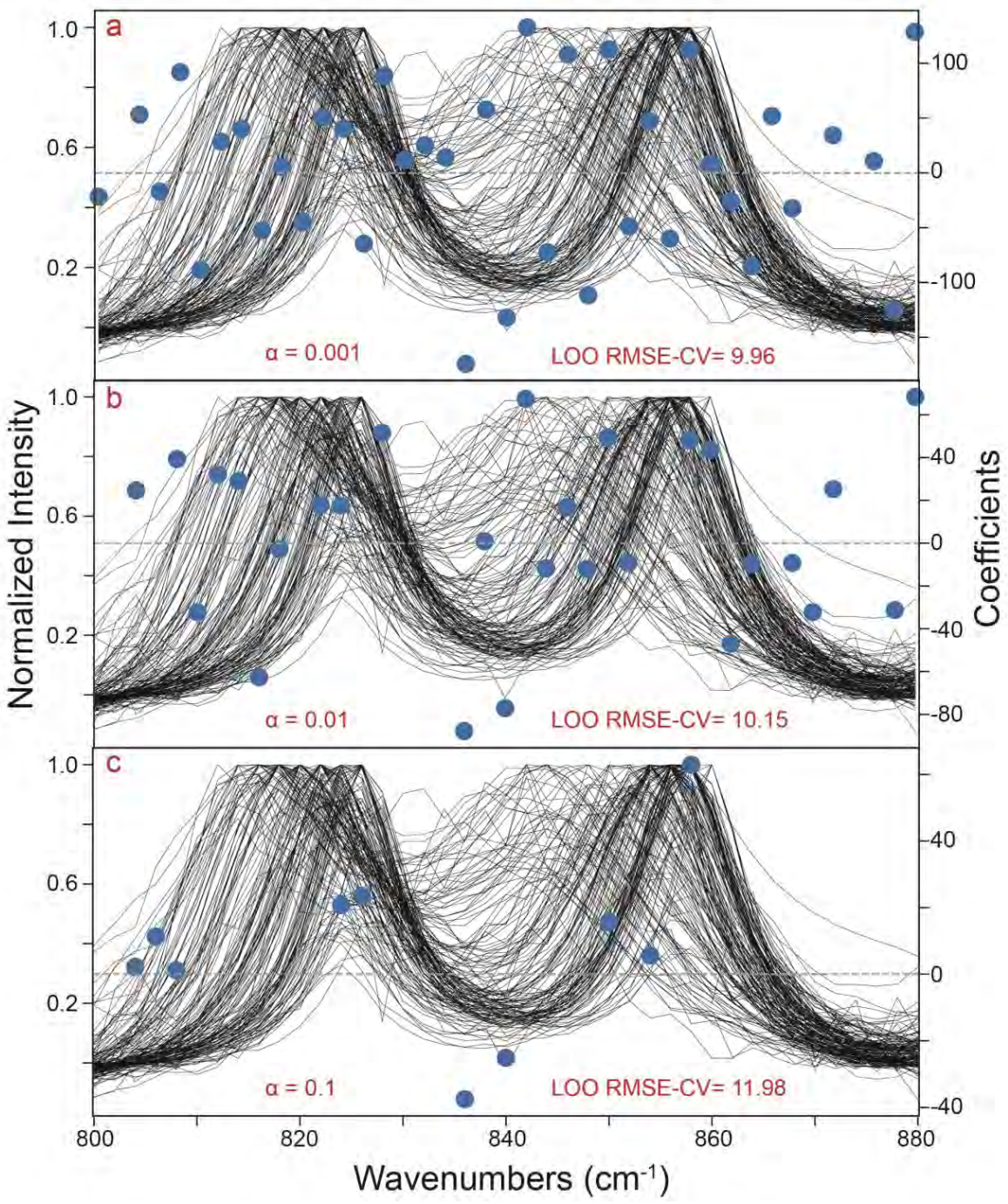


Figure 6. Variations in Lasso model accuracy as a function of the number of coefficients. As α increases, fewer channels are examined: (a) 38 channels for α = 0.001, (b) 27 channels for α = 0.01, and (c) 10 channels for α = 0.1. As the number of channels examined is decreased (fewer coefficients within the model), the RMSE-CV of the models increases in value (gets worse). This demonstrates the value of a models that examines a high number of channels, which is achieved in a small α value Lasso model or PLS models.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

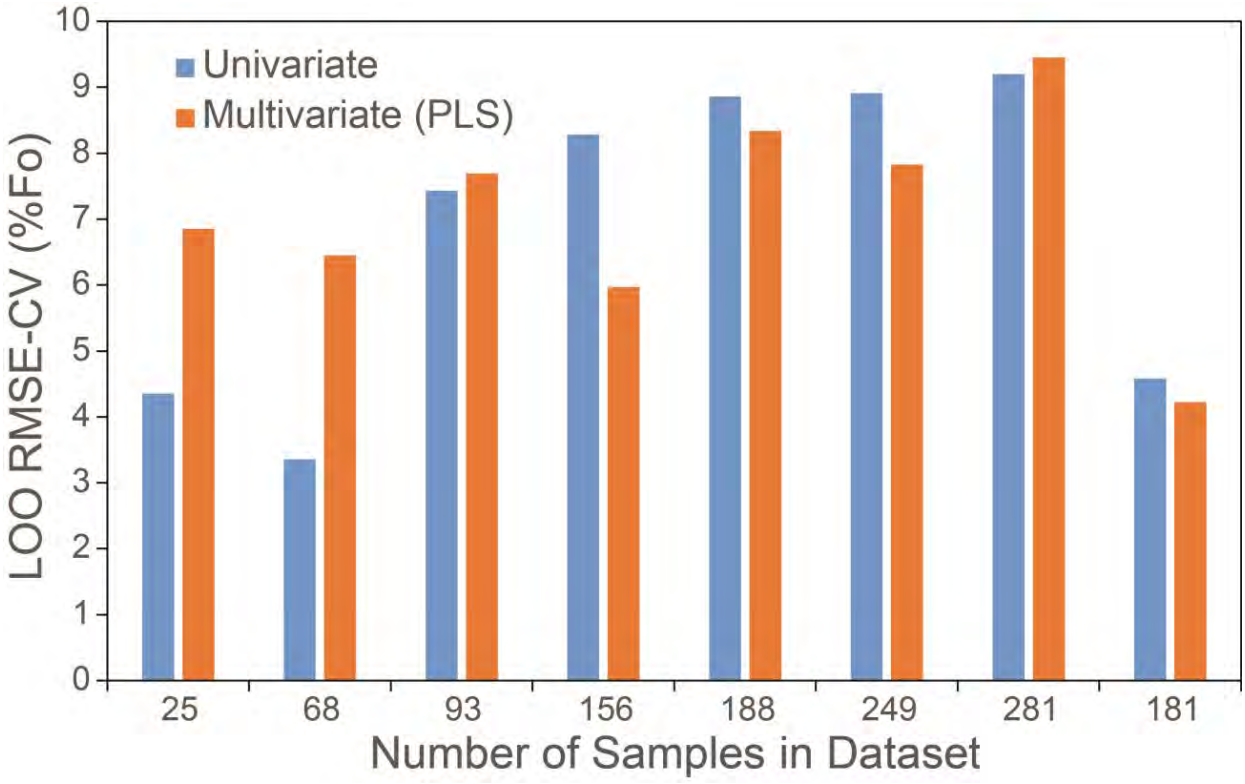
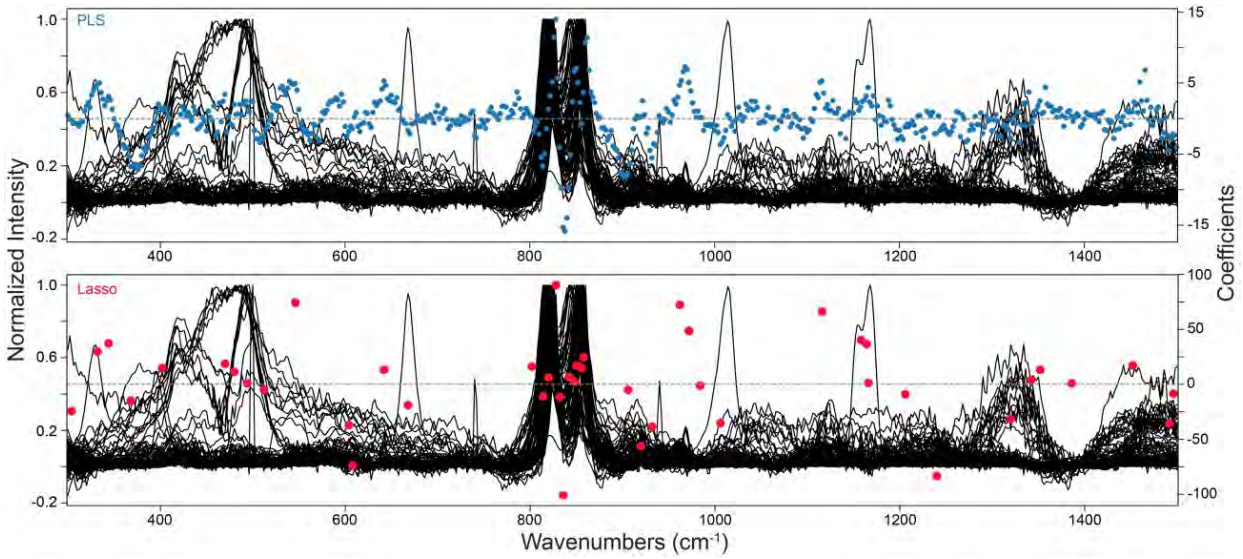


Figure 7. Bar graph comparing results from Tables 4 and 5. When the dataset is small and/or all the data are acquired on the identical instrument, then univariate methods produce better results than those using multivariate analyses. However, as the number of samples and instruments used increase, PLS methods generally produce more accurate results.

This is a preprint, the final version is subject to change, of the American Mineralogist (MSA)
Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2018-6291

47



48

49 Figure 8. (top) Plot of BRAVO and Senterra unaligned data in Table 6, along with circles
50 indicating the magnitude of PLS coefficients (right-hand y axis units). Note that PLS
51 coefficients are proportional to spectral intensity at each wavenumbers, so absolute values cannot
52 be compared on this plot. However, the PLS coefficients do demonstrate that the entire
53 wavenumber range contains information useful to predicting composition. (bottom) Analogous
54 plot for the same data but showing Lasso coefficients for a model with $\alpha = 0.001$.