Comparison of Univariate and Multivariate Models for Prediction of Major and Minor Elements from Laser-Induced Breakdown Spectra with and without Masking

M. Darby Dyar,[a,*] Caleb I. Fassett,[a] Stephen Giguere,[b] Kate Lepore,[a] Sarah Byrne,[a] Thomas Boucher,[b] CJ Carey,[b] and Sridhar Mahadevan[b]

[a]Department of Astronomy, Mount Holyoke College, South Hadley, Massachusetts, 01075, United States

[b]College of Information and Computer Sciences, University of Massachusetts, Amherst, Massachusetts, 01003, United States

[*]Corresponding author at: Department of Astronomy, Mount Holyoke College, South Hadley, Massachusetts, 01075, United States.
E-mail address: mdyar@mtholyoke.edu

ABSTRACT

This study uses 1356 spectra from 452 geologically-diverse samples, the largest suite of LIBS rock spectra ever assembled, to compare the accuracy of elemental predictions in models that use only spectral regions thought to contain peaks arising from the element of interest versus those that use information in the entire spectrum. Results show that for the elements Si, Al, Ti, Fe, Mg, Ca, Na, K, Ni, Mn, Cr, Co, and Zn, univariate predictions based on single emission lines are by far the least accurate, no matter how carefully the region of channels/wavelengths is chosen and despite the prominence of the selected emission lines. An automated iterative algorithm was developed to sweep through all 5485 channels of data and select the single region that produces the optimal prediction accuracy for each element using univariate analysis. For the eight major elements, use of this technique results in a 35% improvement in prediction accuracy; for minors, the improvement is 13%. The best wavelength region choice for any given univariate analysis is likely to be an inherent property of the specific training set that cannot be generalized.

In comparison, multivariate analysis using partial least-squares (PLS) almost universally outperforms univariate analysis. PLS using all the same wavelength regions from the univariate analysis produces results that improve in accuracy by 63% for major elements and 3% for minor element. This difference is likely a reflection of signal to noise ratios, which are far better for major elements than for minor elements, and likely limit their prediction accuracy by any technique. We also compare predictions using specific wavelength ranges for each element against those employing all channels. Masking out channels to focus on emission lines from a specific element occur decreases prediction accuracy for major elements but is useful for minor elements with low signals and proportionally much higher noise; use of PLS rather than univariate analysis is still recommended. Finally, we tested the generalizability of our results by analyzing a second data set from a different instrument. Overall prediction accuracies for the

mixed data sets are higher than for either set alone for all major and minor elements except Ni, Cr, and Co, where results are roughly comparable.

Keywords: Laser-induced breakdown spectroscopy; LIBS; partial least-squares analysis; PLS

1. Introduction

Obtaining quantitative chemical information from laser-induced breakdown spectroscopy (LIBS) spectra of geological samples [1-3] is especially challenging due to textural differences and wide variability in the bulk compositions of naturally occurring glasses and rock-forming minerals. *In situ* analyses [4] using field instruments present further challenges due to sampling conditions (varying distance [5] and incidence angle [6]), especially when done remotely such as on Mars [7,8].

Perhaps the biggest obstacle to quantitative analyses in all complex materials with LIBS is the group of factors known as chemical matrix effects [9,10], which often result in emission peak areas that are not directly proportional to element concentrations. This complicates conventional univariate analyses of individual element peaks in LIBS spectra. Multivariate analyses (MVA) have the potential to alleviate some of the problems introduced by matrix effects by incorporating a broad spectral range rather than a single peak, thereby utilizing all the possible predictive information in each spectrum. Both univariate and MVA have been used in studies of geological materials; in general univariate is employed for quantification of minor elements and MVA is used for majors. However, no systematic investigations have compared univariate and multivariate analyses in large data sets for a wide range of elements and concentrations.

The existing LIBS literature varies widely with respect to uses of these techniques for quantification. Applications can be grouped into those focusing on elements that constitute a significant proportion of the materials (major elements) and on elements with low concentrations

but high interpretive relevance (minor or trace elements). For example, light elements such as Li have been quantified in various types of geological materials including spodumene, tourmaline, and topaz [11-13]. These minor elements are generally quantified using univariate analyses, in which the area or intensity of a single emission peak is related directly to concentration. When the matrix is held constant, as in many industrial applications, this is generally a sound assumption. Even in variable matrices, it is often assumed that chemical matrix effects will have little influence on such small peaks [14]. For example, the ChemCam LIBS team on Mars uses univariate analysis for quantification of some trace elements such as H [15,16], C [17,18], Cr [19], Mn [20,21], and Zn [22], though a modified PLS method is employed for Li, Ba, Sr, and Rb [23].

Several different multivariate approaches have been tested for LIBS applications, including principle components analysis [24] and artificial neural networks [25,26]. Boucher et al.[27] analyzed a geologic data set using linear regression methods including partial least squares (PLS-1 and PLS-2), principal component regression, least absolute shrinkage and selection operator (lasso), elastic net, and linear support vector regression. They compared the linear methods against results from nonlinear regression methods including kernel principal component regression, polynomial kernel support vector regression, and *k*-nearest neighbor regression. Of all these methods, PLS is by far the most common technique used for interpreting LIBS data on geological samples. For example, PLS is currently being used as part of a combined approach with independent components analysis (ICA) to predict major elements compositions of Mars surface materials [28,29]. For minor elements, PLS models may lack interpretability due to the problem of geochemical camouflage, in which trace elements are best predicted by more intense lines from major elements with similar size and charge for which they commonly substitute. This problem can only be overcome by use of artificially-doped standards that break the geochemical substitution [30].

In this study, we compare univariate and multivariate analyses of LIBS data to evaluate

the relative accuracy of these techniques using the largest-ever suite of LIBS data on geological samples assembled, including spectra from 171 doped samples. Our goal is to assess the potential for improvements in accuracy of elemental predictions by testing models that use only spectral regions known to contain peaks arising from the element of interest versus those that use information in the entire spectrum. We test the hypothesis that limiting calibration training set data to element-specific "masked" regions allows element peaks to be more prominently represented in multivariate models, which might improve prediction accuracy. We make direct comparisons between quantification of chemical compositions using individual peaks, groups of individual peaks, and the multivariate analysis technique of partial least squares. We test four permutations of previously-used approaches to elemental prediction accuracy as used for both major and minor elements.

1) Univariate analyses of individual peaks specific to each element, as is typically done in many LIBS applications where only a single matrix is employed;

2) Multivariate (PLS) analysis of the same peaks chosen above for each element, as used successfully by Olilla et al. [23] in predicting Li, Ba, Sr, and Rb;

3) Univariate and multivariate analyses of a wavelength range selected for each element by an automated sweep algorithm; and

4) PLS predictions using unmasked spectra – i.e., the entire spectral region.

Through these analyses, we inform decisions about choice of univariate analysis peaks, when to use PLS versus univariate analysis, and how to construct training sets to yield maximum generalizability of predictive models.

2. Samples and spectral acquisition

Tests performed for this project primarily used a large data set collected at Mount Holyoke College (MHC). The first 280 samples were selected at random from the collections of geological rock powder standards in our laboratory, while the remaining 171 samples came from

a project involving doping of minor elements to create calibration curves. To test the generalizability of conclusions reached on the basis of those data, we also compare our results to a suite of 400 samples (including 106 that are also in the MHC sample suite) for which spectra were acquired at Los Alamos National Laboratory (LANL) as calibrations for the ChemCam instrument on the Mars Science Laboratory rover *Curiosity* [8]. Compositions of all samples studied are represented on a plot of total alkalis vs. silica in **Fig. 1** and characteristics of each data set are given in **Table 1**.

The first set of 840 spectra was acquired on the ChemCam-analog LIBS instrument at MHC, which uses a Quantel Ultra100 laser operating at 1064 nm and up to 20 Hz with a 7-ns pulse width and 3.5-mm beam diameter. A variable attenuator is permanently integrated into the laser, allowing power density to be manipulated by the user to match the range used on Mars. Energy of every pulse is recorded with a Newport 818E series pyroelectric energy detector and meter. Data were recorded under a 7-Torr CO_2 atmosphere over three different energy ranges/spectrometers with spectral resolutions of 0.15-0.25 nm from 220-330 nm, 0.09 nm for 380-470 nm, and 0.42 nm for 490-930 nm. Three laser power densities were selected to bracket the plasma temperatures observed on Mars as indicated by the ratio of intensities of the Si II peak at 634.7 nm to that neutral Si I line at 288.2 nm, which provides a proxy for temperature [31]. There were 280 geologic samples in this suite. The laser powers are designated here as 3.2% (1.9 mJ), 5% (2.8 mJ), and 7% (3.8 mJ), as chosen to reproduce the energy density of ChemCam while compensating for the shorter laser-sample distance at MHC (20.2 cm). Each sample was analyzed in five locations with 30 laser shots to mitigate effects of heterogeneity and the spectra were averaged together to create one spectrum per composition.

Spectra from that sample suite were merged with a suite created for use in calibrating minor elements by LIBS and described in Lepore et al. [30], which evaluates the accuracy of univariate minor element predictions as a function of the composition of the samples' major element matrices and examines factors that limit prediction accuracy of univariate calibrations.

Five different sample matrices were doped with 10-85,000 ppm Cr, Mn, Ni, Zn, and Co and then independently measured in 175 mixtures by XRF, ICP-ES, and LIBS, the latter at three different laser energies (1.9, 2.8, and 3.7 mJ). Univariate prediction models for minor element concentration were created using varying combinations of dopants, matrices, normalization/no normalization, and energy density. Those results show the superiority of using normalization for predictions of minor elements where the predicted sample and those in the training set have matrices with similar SiO_2 contents. Normalization also mitigates differences in spectra arising from laser/sample coupling effects and use of different energy densities. Accordingly, we use normalization in the current study.

For the doped samples from Lepore et al. [30], the same three laser energies, instrumentation, preprocessing sequence, and averaging protocols were used as for the first set of 840 spectra. It includes 172 samples that are mixtures of five bulk rocks (two basalts, a granodiorite, sea sand, and an ultramafic-analog mixture composed of olivine and clinopyroxene) doped with varying amounts of five different trace elements (Ni, Mn, Zn, Cr, and Co). It is important to note that although our doped samples contained in many instances up to ~10 wt.% of the doped element in oxide form, the analyses in this study used only standards with minor element concentrations below 6,000 ppm to better match the concentration levels found in typical rock-forming parageneses on Earth and Mars.

It must also be noted that there are three potentially important differences between the MHC LIBS and the instrument at LANL. First, the LANL data were acquired using stand-off conditions at lower laser power (14 mJ/pulse laser energy from 1.6 m stand-off distance) while the MHC data cover a higher power density (3.8 mJ from ~1.6 cm distance). Second, the LIBS instrument at Mount Holyoke is limited in sensitivity because the spectrometers utilize 1D CCD detectors, making it less sensitive than ChemCam by a factor of (on average) approximately 8 in the UV, 4 in the VIS, and 6 in the VIS/NIR regions. Third, the MHC LIBS is limited by detector

readout noise, so multiple plasmas must be collected in a single spectrometer integration to achieve an acceptable signal-to-noise ratio.

In all analyses at both laboratories, at least 30 shots were acquired on a minimum of five locations on each target, and the resultant >150 spectra were averaged together. The laser beam sizes used were consistently much greater than the grain size in each pressed powder pellet (<1 μm), so this mean spectrum likely is a good representation of the sample even if there is some heterogeneity in the pellets [32]. Thus, overall we utilized a database containing 452 different samples each run at three different laser powers was utilized for the tests in this paper (1356 unique spectra), creating one of the largest known suites of LIBS data for geologic standards available for analysis to date.

Our MHC data were compared against 400 spectra acquired at LANL on the ChemCam Engineering model (flight spare) in a vacuum chamber filled with 7 Torr CO_2 that was placed 1.6 m from the telescope Schmidt plate [31]. Because the laser was optimized to operate under Mars surface conditions (-10 to 0°C), the testbed multiplexer unit was placed in an enclosure and cooled to 4°C, allowing it to achieve the 14 mJ/pulse laser energy used on Mars; all samples were shot using that same laser power. Spectra were collected over three difference ranges, each with its own spectrometer spanning the ultraviolet (240.8-340.8 nm), violet (382.1-469.1 nm) and visible and near infrared (473.2-905.6 nm) regions. Each sample was analyzed in five locations with 50 laser shots per location to mitigate effects of heterogeneity and the spectra were averaged.

3. Prediction models

Univariate (linear least squares) and multivariate (partial least-squares, PLS) analyses of LIBS spectra were used to predict eight major (wt. % SiO_2, Al_2O_3, TiO_2, FeO_T, MgO, CaO, Na_2O, and K_2O) and five trace (Ni, Co, Cr, Zn, Mn reported in ppm) element concentrations with

and without wavelength masks. For all of these samples, concentrations of major and trace elements were measured independently by XRF [33].

Data pre-processing for this project used an adaptation of protocols analogous to those described in Wiens et al. [7] with slight variations. The sequence for ChemCam data processing of lab data includes subtraction of dark spectrum, denoising, continuum removal, wavelength-calibration, multiplication times the instrument response function (IRF), conversion to radiance units, and normalization (as needed), followed by masking (to subtract out regions of high noise in the IRF) in that order. The ChemCam team's continuum removal [7] decomposes each spectrum into a set of cubic spline wavelets, and iteratively finds the local minima in this space to within a user-specified scale, concluding with interpolation of a spline function through the different minima [7].

For both MHC and LANL data, gain curves were used to calculate the instrument response function as described in Wiens et al. [7]. It was noted there that the gain curves change sharply at the edges of each spectrometer's wavelength range. These small changes in intensity result in large changes in calibrated signal, leading to increased noise [8]. Thus in the LANL data, the edges of each spectrometer were masked out of the quantitative analysis models. For parity with LANL and with ChemCam,the MHC protocols also mask the same regions out of the quantitative analysis models, including the channels from 240.811-246.635, 338.457-340.797, 382.138-387.859, 473.184- 492.427, and 849-905.574 nm. Detailed descriptions of these methods and the sample suite are given in Clegg et al. [8].

Our MHC implementation uses a Matlab code to subtract a dark spectrum, perform wavelength calibration using a Ti metal reference run in every carousel, correct for instrument response, normalize to total intensity of each of the three spectral regions, and masking to remove the regions noted above, as required for a particular experiment. Depending on the trial, baseline removal is then applied. For this paper, we used custom baseline removal (Custom BLR) as described in Giguere et al. [34] and Dyar et al. [35]. The method generalizes the

problem of baseline removal by combining operations from previously proposed methods to synthesize new correction algorithms for each application and training set. Custom BLR creates novel methods, discovering new algorithms that maximize the predictive accuracy of the resulting spectroscopic models and yield significant improvements over existing methods [34]. Custom BLR produces significant improvements in prediction accuracy over existing methods across varying geological data sets, instruments, and varying analytical conditions [35].

Optimizations were done separately for each element, and the resultant baseline-corrected spectra (13 sets of them, one for each element of interest) were used for subsequent univariate and multivariate models. PLS and univariate predictions used Python code written for this project. The NumPy package was used for loading and manipulating the spectra, and the Scikit-Learn package was used for all PLS models.

The resulting data were then used to train and evaluate univariate and PLS models. For this, our code uses 5-fold cross validation. To train the univariate models, our code simply regresses the sum intensity in the chosen frequency range on the variable of interest. Training the PLS models requires additional effort because they have a tunable parameter, the number of PLS components used in the model. To tune this parameter, our code uses an inner cross validation loop. Each iteration of this loop randomly splits the training set into sub-training and sub-testing sets. PLS models are trained on the sub-training set, and are then evaluated on the sub-testing set. After completing the inner loops, the number of components is set to the one that resulted in the most accurate models on average. As a result, there is no single value of *n* components that is used in each experiment – it is decided separately for each fold. We also trained our models with a higher upper limit on the number of components (<50); results and analysis are given in the supplement to this paper. Finally, a PLS model is trained on the entire training set using the chosen number of components. For this, our code uses 5-fold cross validation. The accuracy of each model is evaluated using the root mean squared error (RMSE-CV) of cross-validation on the testing set against their true values.

3.1. Masking

We tested three different methods for selecting channels to use in our predictions: channels highly-correlated to the element being predicted, channels selected by an automated sweep routine, and all channels. For the first scheme, we calculated correlation coefficients (r^2) between concentration and the spectral intensity at each wavelength for each element. We chose the top 10-11 peaks with r^2 values greater than 0.9 to become the peak centroids for wavelength masks that were unique to each element as, described in Lepore et al. [30]. The wavelength region around each of the selected channels was broadened using inspection of the data set to include the entire peak and its shoulders, in order to maximize the information used in predictions. Some elements, including K, Mg, Na, and Si did not give rise to many identifiable peaks in our wavelength range. In those cases, we selected as many peaks as possible that were identified as good predictors by the r^2 calculation, being careful to select relevant regions for the element of interest only. **Table 2** lists the chosen regions in order with those yielding highest r^2 values first. Data from these regions were then used in univariate and PLS models to predict each element.

In the second channel-selection scheme, we wrote an iterative algorithm to sweep through all 5485 channels of data and select the single region that produced the lowest RMSE-CV for each element using univariate analysis only (PLS models using this method were too computationally expensive). This procedure facilitates the best possible accuracy for any given univariate procedure; the selected region is designated in Table 2 as "Full Sweep." There was no limit on the wavelength (location) or width of the region chosen, which ranged from a single line (308.788 for Al, 648.89 for Na) to a very narrow region for Ca (from 317.23-317.37 nm, where a Ca II band is located) to a broad region for Zn (458.36-621.88 nm, spanning a broad region where many Zn peaks are located). There was also no requirement for the chosen region to occur near or on a known emission line of the element of interest, though in practice, the chosen

regions always included an appropriate emission line (Table 2). Data from the selected regions were then used in univariate and PLS models to predict each element.

In the third set of comparative models, we used all 5485 channels, and modeled them with PLS only.

We note that there are only philosophical differences between this method for channel and peak selection versus the single-peak approach more generally used by spectroscopists in smaller scale studies. In the traditional approach, peak selection is based on first-principles knowledge of a specific peak (emission line), often the most intense in the spectrum. The peak is generally fit with a Lorentzian, Gaussian, or Voigt (a convolution of Lorentzian and Gaussian) function. Most practitioners acknowledge that in LIBS for geological samples, this convention has many shortfalls: there is no guarantee that the chosen peak won't be too highly overlapped by other peaks from different elements to be useful, it is time-consuming, and the magnitude of the peak is often modified by matrix effects. In our method, we use a completely unbiased selection process to look for the channel that best predicts the variable of interest. By definition, this is a peak (or region of adjoining peaks) that is least likely to be overlapped by those from other elements and/or influenced by matrix effects. Our approach is designed to improve prediction accuracy in large datasets for which it is impractical to use peak fitting or inspect each individual spectra to prevent against using overlapped peaks. It may also produce more accurate results because it makes no assumptions about peak shape. Because we have also optimized baseline removal to improve prediction accuracy, this combination produces the best possible accuracy that can be obtained by predicting a single variable (peak area) against concentration.

Moreover, the practice of summing peak area over the energy range of the peak or peaks determined by the above methods does give comparable peak areas to those obtained using conventional peak fits. Fig. 2 shows two examples that compare fitted Gaussian peak areas for the Al I emission line at 396.3 nm and the Ca II line at 939.4 nm against areas determined by

simply summing the counts under each peak. It is obvious that both approaches produce nearly identical results.

3.2. Model Comparisons

Prediction error results for regression methods are here reported as cross-validated root mean squared errors (RMSE-CV) because these have the same units as the original measurements of sample compositions. In this project, these are either expressed as wt.% oxides for major elements or parts per million for minor elements, in keeping with geochemical conventions. For this purpose, we used K-fold cross-validation, which splits the data set into K approximately equal-sized parts, to train the model and tune its parameters (e.g., the number of components used in partial least squares) before it is tested on a held-out dataset. When models are being fit for a sample in K_i, the other K-1 folds (all K_j folds, $i \neq j$) are used to train the model and the K_i fold is used to test the model.

4. Results and discussion

4.1. Univariate models

Univariate prediction models for all elements and all regions chosen on the basis of high correlation with each specific element are given in **Tables 2** and **3**. Emission lines are listed in the table in order of correlation coefficient for the peak centroid, with highest r^2 first. These results immediately show that for any given element, areas of emission peaks selected on the basis of correlation with a single channel do not provide accurate univariate predictions of the elements with which they are associated. Vagaries in the success of using specific lines here are likely due to overlapping lines from the many other elements present in geological samples. The observed variation makes it apparent that considerable and time-consuming experimentation is clearly necessary to obtain optimal results from individual lines in such complicated matrices.

However, the regions chosen using highly correlated lines are quite interesting in and of themselves, because they are not always the most prominent lines of an element. It is true that in

many cases, the chosen lines are predictable. For example, the strongest K emission lines occur at 766.5 and 769.9 nm, and both those regions are selected by the r^2 method (764.36-768.80 and 768.80-771.77 nm, respectively). On the other hand, many of the lines selected using the high-r^2 criterion are ionized species with relatively modest intensities but they produce more accurate predictions, and the results are sometimes complicated. For example, although the neutral Al line at 396.3 nm is one of the strongest Al lines over the UV-NIR region studied here, use of that line produces a prediction error of ±6.85 wt.% Al_2O_3, while the much smaller Al II and Al III lines ca. 449.1 and 683.9 nm, respectively, produce more accurate analyses with RMSE-CV values of ±6.07 and ±6.10 in units of wt.% Al_2O_3. As a second example, the full sweep algorithm chooses the highly specific single-channel at 648.89 nm as the best predictor of Na, but that region is represented in the NIST database only as lines at 647.63 and 651.42 nm. Finally, a very broad range of wavelengths was selected by the sweep algorithm for Zn. Close inspection of the NIST database shows that it tabulates a small number of Zn lines *from only three sources*; in this case, that compilation may not be considered a comprehensive source for Zn emission characteristics. Thus the sweep may select a broad region where there are unrecognized Zn lines that improve prediction accuracy. There may be many cases where the chosen regions indicate previously un-tabulated emission lines; indeed, our selection process for identification of lines is likely far more rigorous than some of the procedures used in old papers from which the NIST database was compiled.

Although the peaks chosen by the automated sweep algorithm are in many cases surprising, their superiority in predicting elemental concentration cannot be disputed. As seen in Table 2, the sweep method produced considerably better (more accurate) prediction results than any line chosen by the correlation-based regression method, in some cases more than halving (for CaO) the prediction error. For the eight major elements, the average improvement in accuracy obtained by using the sweep method rather than the highest r^2-producing line is 35%; for minors, the improvement is 13%. This suggests that human selection of appropriate peaks or

wavelength ranges for univariate analysis is not necessarily optimal. These results highlight the inherent difficulty of choosing individual peaks by trial and error, and show the superiority of automated routines for selection of peaks for optimal prediction accuracy using univariate analyses.

It is important to note that the best peaks for prediction for any given element found by this study may not be generalizable. They likely vary from sample suite to sample suite because of issues with overlap from other lines in the matrix. Thus, optimal prediction accuracy from univariate analysis not only depends greatly on the choice of emission line, but the best choice may well be an inherent property of the specific training set. Our recommendations for line selections come from a large enough data set that they probably apply to most geological studies. But if the matrix in any given sample suite is different, the sweep procedure should be repeated.

Finally, the brute force method employed in our sweep algorithm is relatively simple and can easily be adapted to any data set. It is a useful way to place a boundary on the best possible accuracy that can be obtained from any univariate prediction(s). Although code for this analysis is available from the authors, we do not endorse its use because, as the following section amply demonstrates, PLS produces consistently better results for major elements on the entire data set, and comparable results for minor elements. It is difficult to imagine a scenario in which univariate analysis would truly outperform PLS, especially in applications with complex matrices.

4.2. PLS models

The difference in prediction accuracy of our limited-range PLS models over their univariate equivalents is shown graphically in a plot that compares their RMSE-CV values in **Fig. 3**. The RMSE-CV numbers are consistently smaller for the PLS models; i.e., they lie below the 1:1 line along which univariate and multivariate models produce identical errors. The PLS models are using the exact same data as in the univariate models, but their performance improvement is impressive. The models benefit from the ability of PLS to exploit the

multicollinearity between the X variables (multiple spectral lines related to the same element or correlated elements) and the elemental abundances (Y response variables). Almost universally, use of PLS produces better prediction accuracy than univariate analysis (Tables 2 and 3) for major elements. This conclusion holds even when the number of channels in any given regression analysis is extremely small (**Fig. 4**). The magnitude of the improvement between univariate and PLS is generally not proportional to the width of the peak interval.

Moreover, the improvement in prediction accuracy for PLS over the univariate approach is dramatic. Prediction accuracy using the Si I peak at 288.2 nm reduces from ±13.76 wt.% SiO_2 for a univariate model to ±8.18 wt% for PLS. The average improvement in RMSE-CV for major elements between univariate and PLS models is 30% for the individual peak models. All of the small-range, single-peak models predicted using PLS out-perform all the univariate models including the sweep algorithm. PLS using all the masked regions (i.e., those listed in Tables 2 and 3) produces results that are 63% more accurate than those for univariate. Given these results, it seems that the use of any type of univariate analysis for prediction of major elements is inadvisable, and would need to be explicitly justified for any given data set.

For minor elements, the difference between univariate and PLS is muted, highest for Mn and least for Zn, with an average improvement of only 3% in prediction accuracy. This difference is likely a function of the magnitudes of the peaks, which are far smaller for minor elements than for major elements. The considerably great signal to noise ratio for these small peaks undoubtedly limits their prediction accuracy by *any* technique.

4.3. Masking to focus on specific element-specific emission lines

All of the models discussed up to this point are masked to focus the predictive analysis on individual regions of the spectrum where a known emission line or lines from the element of interest occur(s). Although these models are thus highly interpretable, they do not take advantage of other lines with potentially different transition probabilities from the same element, nor do

they take matrix effects from other elements into account because their contributions are masked out. Thus, we calculated two additional sets of prediction for models for each element.

In the first set, we used all the energy ranges from Tables 2 and 3 together and employed that total area to predict each element's concentration. Results of this comparison are given in **Table 2** and shown graphically in **Fig. 5**. For major elements, PLS again gave consistently more accurate predictions than univariate, yielding a 63% improvement in prediction accuracy. For minor elements, the relative accuracy improved by 25% overall.

In the second experiment, we used the entire spectral region to predict the elements of interest (**Figs. 5** and **6**). These produced an 11% improvement in accuracy over the all-peaks PLS models for major elements except for FeO_T, which became less accurate. In contrast, prediction accuracy for minor elements got worse by 10% overall when all channels were included in the prediction. Zn actually improved slightly, but all the other minor element predictions became less accurate. These results might be expected because the major elements have a wealth of lines across all spectral regions in our data; masking removes regions of the spectra where there is useful predictive information. Minor elements, however, tend to be swamped by the magnitudes of major elements, so inclusion of the extra channels in those models decreases their effectiveness in multivariate models, though that effect is relatively small. We can conclude from these results that for a single data set acquired using a single instrument (albeit at varying power densities), masking is not needed (indeed, it is deleterious) for prediction of major elements. However, masking is advantageous for prediction of minor elements with low signals and proportionally much higher noise, though use of PLS is still recommended.

4.4. Comparison to ChemCam laboratory calibrations using a different instrument

To investigate the generalizability of our results regarding masking versus not masking for trace and minor elements, we obtained an alternate data set in the form of 400 spectra used by the ChemCam team for calibration as described in Clegg et al. [8] and Anderson et al. [29]. These papers report on new calibrations for the ChemCam instrument using a laboratory LIBS

instrument, Mars-like atmospheric conditions, and standards that span a wider compositional range than previously employed. The new Clegg et al. [8] calibration uses a combination of partial least squares (PLS1) and independent component analysis (ICA) algorithms, together with a calibration transfer matrix to minimize differences between the conditions under which the standards were analyzed in the laboratory and the conditions on Mars. Anderson et al. [29] use the same data set but only PLS to demonstrate a conceptually simple method for improving the accuracy of quantitative LIBS analysis of diverse target materials termed "sub-model" partial least squares. The method is based on training several PLS models on sets of targets with limited composition ranges and then "blending" these "sub-models" into a single final result.

To make this comparison effective, we first compared the prediction accuracies of the models just discussed against those obtained by the ChemCam team on the engineering model (twin of the flight instrument) at Los Alamos National Laboratory (**Fig. 7**). The team data set comprises data acquired using only a single constant laser power density, which ought to make their data set more homogeneous than the larger one examined in the current study. The new ChemCam team model [8] uses a combination of partial least squares (PLS1) and independent component analysis (ICA) algorithms, and employs sub-models to customize the training set for any given prediction by using one of three concentration ranges. An alternative to the team model has also been proposed by Anderson et al. [29] that trains several different PLS models on data sets with small compositional ranges and then blends the resultant sub-models into a single final result. Results from these models are presented in **Table 4**. They show that this smaller, more homogeneous data set from LANL actually does not produce more accurate results than those from the MHC data set.

To explore this conclusion further, we predicted the LANL data set using the same data processing train and PLS code as in the current study. We predicted the ChemCam data set using a single PLS model. There were a few differences in the data sets used. The team identified and removed outliers from their training set using an iterative process unique to each element. They

also applied an Earth-Mars correction factor to all channels before building their models, using a comparison of data from the calibration targets on Mars and in the laboratory to make their calibration suitable for predicting Mars. Finally, the team's approach uses sub-models, in which different models are trained for varying ranges of composition for each element, and then blended back together. Our analysis used custom baseline removal [34, 35] rather than the team's algorithm and our data pre-processing sequence is slightly modified, but it uses all channels of all spectra available. This makes the data set we analyzed rather more heterogeneous than the one actually used by the team.

Results do not support the hypothesis that the smaller LANL data set provides better accuracy than our larger one, even though the MHC model was run on the entire data set, including the outliers discarded by the ChemCam team. In fact, our simpler model (one PLS model trained on all channels of data) provides comparable to or better accuracy than the team's methods [8, 29] for major element predictions. We speculate that models built using the MHC data set, which intentionally spans multiple laser power densities, can better accommodate variations that result from pulse-to-pulse signal fluctuations typically found in LIBS instruments.

Interestingly, our minor element models for the LANL data do give significantly better accuracy than those from the MHC data set (**Fig. 8**), but this might be expected for two reasons. First, all the LANL data were acquired from naturally occurring samples in which concentrations of the minor elements are low, in contrast with the MHC data set in which samples are doped with up to 10 wt.% of the "minor" elements. Second, our previous work has documented a tendency for multivariate analyses of minor elements to employ the emission lines for major elements to which they are related. This is the effect of geochemical camouflage, which occurs when the radii between major and minor elements differ by less than 15%, minor element cations can substitute into mineral sites typically occupied by major elements with the same charge [36]. For example, Rb^{1+} can be predicted using K^{1+} emission lines because that is a common substitution in feldspar minerals. The 172 doped samples in the MHC data subvert those

correlations with major elements. Thus, predictions using the MHC data would be expected to show prediction accuracies based only on actual emission lines from each specific minor elements rather than lines caused by camouflaging major elements. Clearly, accuracy receives a boost when the major element lines can be used in predictions, and for terrestrial studies, such camouflage should be exploited. However, for Mars predictions and for the sake of interpretability, it may be advantageous to use doped samples.

Finally, these results provide points of comparison for masked and unmasked models employing a hybrid data set that combines the Los Alamos and MHC data suites, as follows.

4.5. Combining data sets from different instruments together

As a final test, we investigated the dependence of prediction accuracy on the combined data set with a total of 1756 spectra. We make this comparison to illustrate the mingled effects of three factors on the question of masking versus not masking: spectrometer differences, training set size, and single vs. multiple power densities. Results are shown in Table 4 and Figs. 7 and 8 for use of both data sets employing the masks from Tables 2 and 3 and without masks. We note that baseline removal was again customized as part of the prediction algorithm.

Overall, absolute prediction accuracies for the mixed data sets are generally higher (worse) then for either set alone for most major and minor elements, despite the expectation that a larger training set would produce more accurate predictions. Recent work by Thomas Boucher in our group suggests that this occurs when merging data sets from very different instruments or collected using dissimilar analytical conditions. Those criteria are certainly true in this situation: Mount Holyoke data were collected at a 20.2 cm standoff distance with multiple power densities (3.2, 2.8, and 3.8 mJ), while LANL results were acquired from a distance of 1.6 m at a single power density (14 mJ). When analyzing these data sets together, PLS cannot reconcile those differences, and worse prediction accuracy results. This problem can be solved by use of calibration transfer techniques that correct systematic differences between data sets and look for the most predictive commonalities between them. Those experiments are outside the scope of the

current work. However, if we were to repeat this analysis using a calibration transfer method (like the correlation analysis domain adaptation we are developing) to first align the results, we would expect that the combination data set would perform better than either one alone.

5. Implications and conclusions

For this data set (the largest suite of LIBS rock spectra ever assembled) and these elements, univariate predictions based on single emission lines are by far the least accurate, no matter how carefully the region of channels/wavelengths is chosen. This result is expected given the wide range of matrix compositions in our geologically-relevant calibration suite, and is likely generalizable to any system with similarly variable matrices. The traditional method of subjective choice of elemental lines for quantitative analyses in complex systems does not produce the best possible results. Univariate analyses based on use of a computationally-optimized wavelength range for each element (here referred to as the sweep algorithm) show improved accuracy over univariate predictions based on individual peaks in all cases, and place a bound on the accuracy obtainable using human-chosen ranges. These results support the general conclusion that users of univariate analysis for elemental quantification would do well to employ an optimization approach to the choice of peaks. Significant improvements in prediction accuracy may be anticipated as a result, especially in samples with complex matrices containing many different elements. The advantages of the optimization approach over human selection should apply to all LIBS applications. Source code for this procedure is available from the corresponding author.

However, univariate analysis is <u>not</u> the recommended technique for prediction of major elements in geological samples. In nearly every single test performed in this study, PLS outperformed univariate analysis in prediction accuracy. This result is expected for three reasons. First, elements that are present in the high concentrations will have many different emission lines throughout the spectral region studied. Any type of analysis that excludes regions of the

spectrum where there is possibly predictive information will, by definition, denigrate prediction accuracy. Second, univariate analyses may be subject to problems from overlap in anything but a pure matrix. In complex training sets with compositional diversity, it can be extremely difficult to anticipate by trial and error where those overlaps will occur. Finally, matrix effects arise from the ensemble of elements that constitute any given plasma. Thus, there is useful data not only in the emission lines of an element of interest, but also in those arising from other elements. This effect was documented in Table 2 of Dyar et al. [36], where LIBS prediction accuracy was evaluated with the least absolute shrinkage and selection operator (lasso), a penalized shrunken regression method that selects the specific channels for each element that explain the most variance in the concentration of that element. Specific lines chosen for their predictive usefulness by the lasso were in no case solely those from the element being predicted. Some of the emission lines came from other elements with known geochemical correlations or close overlap of emission lines, but many resulted from unrelated elements and thus must arise from matrix effects caused by an underlying physical process. The best possible prediction accuracy therefore comes when the PLS analysis has access to channels where emission lines of other peaks occur. For all these reasons, univariate analysis of LIBS lines should not be the first choice for quantitative analysis of major elements in any application, unless it is clear that these three problems do not apply.

The same general conclusion of PLS superiority over univariate can also be made for minor elements, though improvements in prediction accuracy when changing from optimally masked spectra regions to full-spectral PLS are more nuanced. Our results support the conclusion of Ollila et al. [23], who used a small training set with un-doped samples to demonstrate that PLS over a limited wavelength range yielded only slightly improved accuracy in predictions of Rb and Sr but not for Li and Ba. However, compositional diversity in the matrix does still affect quantitative analysis of minor elements. Major element features can act to obscure any extra information on minor elements that may be present in the rest of the spectrum. While not a

matrix effect per se, the number and magnitude of emission lines from major elements in any sample can also have dramatic effects on the magnitudes of minor element peaks in a LIBS spectrum when normalization is used to correct for changes in the amount of material ablated [30]. Overall, probably the biggest limitation to prediction of minor elements with LIBS is simply the low signal to noise ratio of their emission lines, which may be the reason why the prediction accuracies for trace elements are generally poor overall (especially when considered as percentages of the total concentration present, c.f. Fig. 8). Thus use of limited wavelength ranges for PLS and univariate analyses for minor element predictions both employ the same essential set of noisy information. Only subtle advantages of one over the other may result, depending on the vagaries of signal to noise and the presence of overlapping peaks.

Acknowledgments

This work was supported by NSF grants CHE-1306133 and CHE-1307179 and NASA grants NNX14AG56G, NNX12AK84G, and NNX15AC82G. Student support from the Massachusetts Space Grant Consortium is also gratefully acknowledged. We thank two anonymous reviewers for extremely helpful comments that substantially improved this paper.

References

[1] P. Pořízka, A. Demidov, J. Kaiser, J. Keivanian, L. Gornushkin, U. Panne, J. Riedel, Laser-induced breakdown spectroscopy for in situ qualitative and quantitative analysis of mineral ores, Spectrochim. Acta Part B 101 (2014) 155-163.

[2] N.J. McMillan, R.S. Harmon, F.C. DeLucia, A.W. Miziolek, Laser-induced breakdown spectroscopy analysis of minerals: carbonate and silicates, Spectrochim. Acta B. 62 (2007) 1528-1536.

[3] R.S. Harmon, J. Remus, N.J. McMillan, C. McManus, L. Collins, J.L. Gottfried Jr., F.C. DeLucia, A.W. Miziolek, LIBS analysis of geomaterials: geochemical fingerprint for the rapid analysis and discrimination of minerals, Appl. Geochem. 24 (2009) 1125–1141.

[4] R.T. Wainer, R.S. Harmon, A.W. Miziolek, K.I. McNesky, P.D. French, Analysis of environmental lead contamination: comparison of LIBS field and laboratory instruments, Spectrochim. Acta B. 56 (2001) 777-793.

[5] R.C. Wiens, S. Sharma, J. Thompson, A. Misra, and P.G. Lucey, Joint analyses by laser-induced breakdown spectroscopy (LIBS) and Raman spectroscopy at stand-off distances, Spectrochim. Acta B. 61 (2005) 2324-2334.

[6] E.A. Breves, K.H. Lepore, M.D. Dyar, Laser-induced breakdown spectroscopy of glasses and rocks at varying ablation and collection angles, Lunar Planet. Sci. Conf. 47 (2015) Conf. Abstract #2536.

[7] R.C. Wiens, S. Maurice, J. Lasue, O. Forni, R.B. Anderson, S. Clegg, S. Bender, D. Blaney, B.L. Barraclough, A. Cousin, L. Deflores, D. Delapp, M.D. Dyar, C. Fabre, O. gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P.-Y. Meslin, H. Newsom, A. Ollila, R. Perez, R.L. Tokar, D. Vaniman, Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover, Spectrochim. Acta B. 82 (2013) 1-28.

[8] S.M. Clegg, R.C. Wiens, R. Anderson, O. Forni, J. Frydenvang, J. Lasue, A. Pilleri, V. Payré, T. Boucher, M.D. Dyar, S.M. McLennan, R.V. Morris, T.G. Graff, S.A. Mertzman, B.L. Ehlmann, S.C. Bender, R.L. Tokar, I. Belgacem, H. Newsom, B.C. Clark, M. Melikichi, A. Mezzacappa, R.E. McInroy, R. Martinez, P. Gasda, O. Gasnault, S. Maurice, Recalibration of the Mars Science Laboratory ChemCam instrument with an expanded geochemical database, Spectrochim. Acta B, submitted.

[9] A. Segnini, A.A. Pereira Xavier, P.L. Otaviani-Junior, E. C. Ferreira, A. M. Watanabe, M.A. Sperança, G. Nicolodelli, P.R. Villas-Boas, P.P.A. Oliveira, D.M.B. Pereira Milori,

Physical and chemical matrix effects in soil carbon quantification using laser-induced breakdown spectroscopy, Am. J. Anal. Chem. 5 (2014) 722-729.

[10] J.M. Tucker, M.D. Dyar, M.W. Schaefer, S.M. Clegg, S.M., R.C. Wiens, R.C, Optimization of laser-induced breakdown spectroscopy for rapid geochemical analysis, Chem. Geol. 277 (2010) 137-148.

[11] C. Fabre, M.-C. Boiron, J. Dubessy, A. Chabiron, B. Charoy, T. Martin Crespo, Advances in lithium analysis in solids by means of laser-induced breakdown spectroscopy: an exploratory study, Geochim. Cosmochim. Acta. 66 (2002) 1401–1407.

[12] M. Rossi, M. Dell'Aglio, A. De Giacomo, R. Gaudiuso, G.S. Senesi, O. De Pascale, F. Capitelli, F. Nestola, M.R. Ghiara, Multi-methodological investigation of kunzite, hiddenite, alexandrite, elbaite and topaz, based on laser-induced breakdown spectroscopy and conventional analytical techniques for supporting mineralogical characterization, Phys. Chem. Mins. 41 (2014) 127–140.

[13] M.T. Sweetapple and S. Tassios, Laser-induced breakdown spectroscopy (LIBS) as a tool for in situ mapping and textural interpretation of lithium in pegmatite minerals, Amer. Mineral. 100 (2015) 2141-2151.

[14] R. Wang, X, Ma, Q. Yu, Y. Song, H. Zhau, M. Zghang, Y. Liao, Methods of data processing for trace elements analysis using laser induced breakdown spectroscopy, Plasma Sci. Techn. 17 (2015) 944-947.

[15] S. Schröder, P.-Y. Meslin, O. Gasnault, S. Maurice, A. Cousin, R.C. Wiens, W. Rapin, M.D. Dyar, N. Mangold, O. Forni, M. Nachon, S. Clegg, J.R. Johnson, J. Lasue, S. Le Mouélic, A. Ollila, P. Pinet, V. Sautter, D. Vaniman, Hydrogen detection with ChemCam at Gale crater, Icarus 249 (2015) 43-61.

[16] P.-Y. Meslin, O. Gasnault, O. Forni, S. Schröder, A. Cousin, G. Berger, S.M. Clegg, J. Lasue, S. Maurice, V. Sautter, and the MSL Science Team, Soil diversity and hydration

as observed by ChemCam at Gale Crater, Mars, Science 341 (2013) DOI: 10.1126/science.1238670.

[17] J. Lasue, S. Maurice, A. Cousin, O. Forni, P.-Y. Meslin, W. Rapin, S. Schröder, A. Ollila, G. Berger, N. Bridges, S.M. Clegg, C. d'Uston, C. Fabre, O. Gasnault, W. Goetz, J. Johnson, N. Lanza, S. Le Mouélic, M.B. Madsen, N. Mangold, N. Melikechi, A. Mezzacappa, H. Newsom, R.C. Wiens, MSL Science Team, ChemCam analysis of martian fine dust, Lunar Planet. Sci. 45 (2014) Conf. Abstract #1777.

[18] P. Beck, O, Forni, J. Lasue, E. Lewin, A. Cousin, S. Maurice, P.-Y. Meslin, W. Rapin, O. Gasnbault, R.C. Wiens, N. Mangold, V. Sautter, P. Coll, C. Szopa, T. Dequaire, J.G. Blank, and the MSL Science Team, Carbon detection with ChemCam: Laboratory studie and Mars results, Lunar Planet. Sci. 47 (2015) Conf. Abstract #1826.

[19] C. Fabre, A. Cousin, R.C. Wiens, A. Ollila, O. Gasnault, S. Maurice, V. Sautter, O. Forni, J. Lasue, R. Tokar, D. Vaniman, N. Melikechi, In situ calibration using univariate analyses based on the on-board ChemCam targets: first prediction of Martian rock and soil compositions, Spectrochim. Acta B. 99 (2014) 34-51.

[20] A. Cousin, O. Forni, S. Maurice, O. Gasnault, C. Fabre, V. Sautter, R.C. Wiens, J. Mazoyer. Laser induced breakdown spectroscopy library for the Martian environment, Spectrochim. Acta B. 66 (2011) 805-814.

[21] N.L. Lanza, W.W. Fischer, R.C. Wiens, J. Grotzinger, A. Ollila, A. Cousin, R. Anderson, B.C. Clark, R. Gellert, N. Mangold, S. Maurice, S. Le Mouélic, M. Nachon, M. Schmidt, J. Berger, S. Clegg, O. Forni, C. Hardgrove, N. Melikechi, H. Newsom, V. Sautter, High manganese concentrations in rocks at Gale crater, Mars, Geophys. Res. Letts. 41 (2014) 5755-5763.

[22] J. Lasue, S.M. Clegg, O. Forni, A. Cousin, R.C. Wiens, N. Lanza, N. Mangold,L. LeDeit, O. Gasnault, S. Maurice, J.A. Berger, K. Stack, D. Blaney, C. Fabre, W. Goetz, J. Johnson, S. Le Mouélic, M. Nachon, V. Payré, W. Rapin, D.Y. Sumner, Observation of

>5 wt% zinc by ChemCam LIBS at the Kimberly, Gale Crater, Mars, J. Geophys. Res. Submitted.

[23] A.M. Ollila, H.E. Newsom, B. Clark III, R.C. Wiens, A. Cousin, J.G. Blank, N. Mangold, V. Sautter, S. Maurice, S.M. Clegg, O. Gasnault, O. Forni, R. Tokar, E. Lewin, M.D. Dyar, J. Lasue, R. Anderson, S.M. McLennan, J. Bridges, D. Vaniman, N. Lanza, C. Fabre, N. Melikechi, G.M. Perrett, J.L. Campbell, P.L. King, B. Barraclough, D. Delapp, S. Johnstone, P.-E. Meslin, A. Rosen-Gooding, J. Williams, and the MSl Science Team, Trace element geochemistry (Li, Ba, Sr, and Rb) using Curiosity's ChemCam: Early Results for Gale Crater from Bradbury Landing Site to Rocknest, JGR Planets 119 (2014) DOI: 10.1002/2013JE004517.

[24] J.B. Sirven, J.B., B. Salle, P. Mauchien, J.L. Lacour, S. Maurice, G. Manhes, Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods, J. Anal. At. Spectrom. 22 (2007) 1471–1480.

[25] R.B. Anderson, R.V. Morris, S.M. Clegg, J.F. Bell III, R.C. Wiens, S.D. Humphries, S.A. Mertzman, T.G. Graff, R. McInroy, R, The influence of multivariate analysis methods and target grain size on the accuracy of remote quantitative chemical analysis of rocks using laser induced breakdown spectroscopy, Icarus 215 (2011) 608–627.

[26] F. Anabitarte, A. Cobo, J.M. Lopez-Higuera, Laser-Induced Breakdown Spectroscopy: fundamentals, applications, and challenges, ISRN Spectrosc. (2012) http://dx.doi.org/10.5402/2012/285240 (Article ID 285240).

[27] T.F. Boucher, M.V. Ozanne, M.L. Carmosino, M.D. Dyar, S. Mahadevan, E.A. Breves, K.H. Lepore, S.M. Clegg, Nine machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy, Spectrochim. Acta B. 107 (2015) 1-10.

[28] R.C. Wiens, S. Maurice, J. Lasue, O. Forni, R.B. Anderson, S. Clegg, S. Bender, D. Blaney, B.L. Barraclough, A. Cousin, L. Deflores, D. Delapp, M.D. Dyar, C. Fabre, O. gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P.-Y. Meslin, H. Newsom, A. Ollila, R. Perez, R.L. Tokar, D. Vaniman, Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover, Spectrochim. Acta B. 82 (2013) 1-28.

[29] R.B. Anderson, S.M. Clegg, J. Frydenvang, R.C. Wiens, S. McLennan, R.V. Morris, B. Ehlmann, M.D. Dyar, T. Boucher, Improved accuracy in quantitative laser-induced breakdown spectroscopy using sub-model partial least squares, Spectrochim. Acta B. Submitted.

[30] K. Lepore, L.B. Breitenfeld, M.N. Ketley, A.L. Roberts, M.D. Dyar, C.I. Fassett, E.C. Sklute, G.J. Marchand, J.M. Rhodes, M. Vollinger, S.A. Byrne, M.C. Crowley, T.F. Boucher, S. Mahadevan, Cr, Ni, Mn, Co, and Zn calibrations for use in laser-induced breakdown spectroscopy studies of geological materials, Appl. Spectr., submitted.

[31] R.L. Tokar, R.C. Wiens, S. Maurice, A. Pilleri, R. Gellert, R.B. Anderson, S.C. Bender, S.M. Clegg, M.D. Dyar, C. Fabre, O. Forni, O. Gasnault, J. Lasue, N. Melikechi, Relationship between MSL/ChemCam laser focus, plasma temperature, and compositional calibrations, Lunar Planet. Sci. XLVI, 2015, #1369 (abstr.).

[32] M. McCanta, P.A. Dobosh, M.D. Dyar, Testing the veracity of LIBS analyses on Mars using the LIBSSIM program, Space Science Reviews 81 (2013) 48-54.

[33] J.M. Rhodes, M.J. Vollinger, Composition of basaltic lavas sampled by phase-2 of the Hawaii Scientific Drilling Project: Geochemical stratigraphy and magma types, Geochem. Geophys. Geosyst. 5 (2004) Q03G13, doi:10.1029/2002GC000434.

[34] S. Giguere, T. Boucher, CJ Carey, S. Mahadevan, M.D. Dyar, A fully-customized baseline removal framework for spectroscopic applications, Appl. Spec., submitted.

[35]M.D. Dyar, S. Giguere, CJ Carey, T. Boucher, S. Mahadevan, Comparison of baseline removal methods for laser-induced breakdown spectroscopy of geological samples. Spectrochim. Acta B, submitted.

[36] M.D. Dyar, M.L. Carmosino, E.A. Breves, M.V. Ozanne, S.M. Clegg, R.C. Wiens, Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples, Spectrochim. Acta B 70 (2012) 51-67.

[37] M.D. Dyar, M.L. Carmosino, J.M. Tucker, E.A. Brown, S.M. Clegg, R.C. Wiens, J.E. Barefield, J.S. Delaney, G.M. Ashley, S.G. Driese, Remote laser-induced breakdown spectroscopy analysis of East African Rift sedimentary samples under Mars conditions, Chem. Geol., 294-295 (2012) 135-151.

[38] B.-H.Mevik, R. Wehrens. The pls package: principal component and partial least squares regression in R, J. Statist. Software 18 (2007) 1–24.

Table 1. Characteristics of Data Sets Used*

	MHC Data Set				LANL Data Set				Both
	Minimum	Maximum	Average	S.D.	Minimum	Maximum	Average	S.D.	Average
SiO_2	0.05	99.93	57.20	16.65	0.00	98.00	53.9	17.31	56.44
Al_2O_3	0.01	41.80	12.08	6.87	0.00	38.79	14.29	6.99	12.59
TiO_2	0.00	7.35	1.21	1.40	0.00	3.59	0.91	0.82	1.14
FeO_T	0.00	57.50	7.20	5.22	0.00	86.28	7.62	7.78	7.30
MgO	0.00	47.37	7.13	9.55	0.00	56.14	4.65	6.92	6.56
CaO	0.00	50.30	6.02	6.93	0.00	56.42	6.09	8.74	6.03
Na_2O	0.00	5.91	2.00	1.42	0.00	43.97	2.29	2.92	2.06
K_2O	0.00	7.72	1.82	1.77	0.00	12.11	2.19	1.92	1.91
Ni	0	5548	291	645	0	2782	103	269	260
Mn	0	4077	153	443	5	2200	106	228	147
Zn	0	5809	1057	841	0	4879	731	704	1015
Cr	0	5995	256	641	1	2908	186	330	245
Co	0	4428	125	493	0	259	26	30	110

*Units are in weight percent for the oxides and parts per million for the minor elements.

Table 2. Wavelength Regions Used for Masking and RMSE-CV Values for Masked Major Element Data*

	Range (nm)	Peak and Energy (nm)	Univariate RMSE-CV	PLS RMSE-CV	Range (nm)	Peak and Energy (nm)	Univariate RMSE-CV	PLS RMSE-CV
SiO_2	287.90-288.74	Si I 288.2	13.76	8.18	504.99-508.18	Si II 504.1, 505.6	16.40	16.27
	632.42-639.23	Si II 637.1	15.44	11.89	408.31-409.59	Si IV 409.0	16.37	13.35
	456.63-457.18	Si III 456.9	15.48	14.35	390.36-390.97	Si I 390.5	14.04	10.77
	412.45-414.08	Si II 413.1	15.75	9.77	**469.05-474.32**	**Si I, III many**	**8.66**	
Al_2O_3	308.12-308.55	Al I 308.3	6.85	5.43	702.20-707.50	Al II 704.4, 705.9	6.03	6.57
	309.13-309.68	Al I 309.4	6.79	5.30	451.04-451.77	Al II 449.1	6.07	4.80
	395.87-396.57	Al I 396.3	6.85	5.94	465.99-467.13	Al II 466.8	6.89	5.43
	621.64-625.25	Al II 624.5	6.88	4.94	683.32-684.56	Al II 683.9	6.10	5.96
	394.14-394.85	Al I 394.5	6.76	3.47	568.14-570.76	Al III 569.8	6.37	5.34
					308.787	**Al II 308.9**	**4.93**	
TiO_2	323.35-324.65	Ti II 323.5-324.3	1.08	0.74	325.10-325.70	Ti II 325.4, 325.5	1.21	0.89
	334.64-335.44	Ti II 334.7-335.0	1.06	0.74	307.79-308.12	Ti II 308.0	1.37	1.30
	335.92-336.64	Ti I, II 336.2	1.34	1.54	324.65-325.10	Ti II 325.0	1.02	0.91
	336.83-337.70	Ti II 337.1-337.4	1.16	0.67	332.05-332.62	TI II 332.3, 332.4	1.40	1.01
	333.92-334.59	Ti II 334.1, 334.2	1.03	0.95	322.72-323.14	TI II 323.0	1.11	1.01
					336.68-337.35	**Ti II many**	**0.81**	
FeO_T	246.69-277.41	Fe II many	4.84	2.41	388.47-389.29	Fe I 388.7-388.8	4.99	4.63
	283.04-305.08	Fe I-III many	5.12	2.80	404.19-441.15	Fe I 404.7	4.24	2.21
					263.10-263.25	**Fe II 263.2**	**3.17**	
MgO	292.62-293.27	Mg I 293.7	4.15	4.75	515.39-521.68	Mg I 517.4-518.5	9.47	7.93
	446.09-450.79	Mg II 448.2, 448.3	5.94	3.39	278.54-281.15	Mg II 279.2	5.76	3.28
	293.44-294.16	Mg I 293.8	5.97	5.27	284.99-285.60	Mg I 285.3	6.41	5.29
	786.11-793.42	Mg II 787.9, 789.9	9.53	6.91	**445.72-448.12**	**Mg II, III, V many**	**3.03**	
CaO	445.31-446.28	Ca I 445.7	6.80	6.29	317.26-318.98	Ca II 318.0	3.27	3.33
	611.05-613.23	Ca I 612.4	6.89	4.51	392.00-394.20	Ca II 393.4	6.75	5.45
	443.17-444.19	Ca I many	5.42	2.55	315.67-316.46	Ca II 316.0	6.88	3.64
	640.99-651.55	Ca I many	5.10	2.66	442.36-443.06	Ca I many	5.69	2.78
	731.25-733.38	Ca I 732.8	4.93	5.38	396.49-397.35	Ca II 397.0	6.65	6.42
	713.02-715.53	Ca I 715.0	4.99	4.82	**317.23-317.37**	**Ca II 317.9**	**2.39**	
Na_2O	586.59-590.74	Na I 589.2, 689.8	1.39	1.24	312.80-313.18	Na II 312.6	1.42	1.36
	816.76-821.37	Na I 818.6	1.41	1.22	**648.89**	**Na II 647.5**	**0.98**	
K_2O	764.36-768.80	K I 766.7	1.74	1.52	690.05-691.65	K I 691.3	1.77	1.75
	768.80-771.77	K I 766.7	1.47	1.45	454.73-456.23	K II 460.0	1.77	1.46
	692.92-694.85	K I 694.1	1.73	1.65	**824.24-860.22**	**K I many**	**1.20**	

*Range for full sweep algorithm results given in bold face and shaded cells. The average number of components and standard deviation on components for each model across folds are given in supplementary document Table 2S.

Table 3. Wavelength Regions Used for Masking and Univariate and PLS RMSE-CV Values for Masked Minor Element Data*

	Range (nm)	Peak and Energy (nm)	Univariate RMSE-CV	PLS RMSE-CV	Range (nm)	Peak and Energy (nm)	Univariate RMSE-CV	PLS RMSE-CV
Ni	301.15-301.44	Ni I 301.3	599	558	305.37-305.68	Ni I 305.5	602	573
	300.19-300.58	Ni I 300.3	596	536	739.23-739.88	Ni I 739.6	622	616
	440.12-440.33	Ni I 440.3	572	546	741.94-742.78	Ni I 742.4	626	626
	313.31-313.77	Ni I 313.5	609	557	303.65-304.07	Ni I 303.9	593	579
	761.49-762.50	Ni I 761.9	620	626	305.03-305.37	Ni I 305.2	600	564
					261.03-261.08	**Ni II 261.0**	**508**	
Mn	403.27-403.79	Mn I 403.4	814	708	267.06-267.63	Mn II 267.3	850	848
	270.03-270.39	Mn II 270.2	851	810	294.77-295.26	Mn II 295.0	780	693
	288.89-289.14	Mn II 288.8	848	804	600.75-602.94	Mn I 601.5	839	826
	293.84-294.14	Mn II 294.0	840	774	293.21-293.50	Mn II 293.4	835	803
	403.00-403.27	Mn I 403.2	814	684	404.04-404.36	Mn I 404.3	818	723
					260.92-261.08	**Mn II 260.6, 261.0**	**832**	
Zn	330.18-331.78	Zn I 330.4	440	443	328.11-328.43	Zn I 328.3	426	441
	467.99-468.50	Zn I 468.1	432	439	758.37-759.63	Zn II 759.0	436	430
	635.52-636.60	Zn I 636.4	437	443	250.08-250.55	Zn II 250.3	435	436
	255.71-256.18	Zn II 255.9	433	440	773.05-773.89	Zn II 773.5	431	443
	334.41-334.81	Zn I 334.7	427	442	609.27-610.79	Zn II 610.4	415	441
					458.36-621.88	**Zn I, II, many**	**374**	
Cr	519.40-521.42	Cr I 520.2, 520.7	637	641	425.43-425.79	Cr I 425.6	600	478
	283.39-283.84	Cr II 283.6	615	565	313.12-313.53	Cr II 313.3	619	635
	427.45-427.79	Cr I 427.6	631	626	276.53-277.12	Cr II 276.7	626	641
	267.47-268.13	Cr II 268.0	643	655	312.41-312.84	Cr II 312.6	635	625
	284.24-284.54	Cr II 284.4	640	526	301.29-301.68	Cr I 301.5	611	635
					437.82-553.24	**Cr I, II many**	**555**	
Co	257.95-258.49	Co II 258.3	456	411	389.23-389.80	Co I 389.5	484	483
	399.50-399.74	Co I 399.6	458	483	266.24-266.61	Co II 266.4	485	488
	533.47-534.59	Ci II 533.5	474	487	269.47-269.73	Ci II 269.5	475	492
	255.87-256.18	Co II 256.1	463	487	411.85-412.39	Co II 412.2	486	474
	304.32-304.65		485	490	270.59-271.10	Co II 270.8	476	488
					258.27	**Co II 258.2-258.3**	**418**	

*Range for full sweep algorithm results given in bold face and in shaded cells. The average number of components and standard deviation on components across folds for each model are given in supplementary document Table 3S.

Table 4. RMSE-CV Values for Masked Models Compared with Unmasked Models*

Element	MHC Data (this paper)				Los Alamos Data: 400 Standards**					Modeled Together	
	Full Sweep Univariate	All Masks Univariate	All Masks PLS	No Mask PLS	ChemCam Team ICA Model [8]	Anderson Full PLS Model[24]	Anderson Blended Sub-Model	ChemCam Team Model[8]	Our Model No Mask	All Masks PLS	No Masks
Wt.% SiO_2	8.66	14.22	5.23	4.69	8.31	5.66	4.91	5.83	5.18	6.31	5.59
Wt.% Al_2O_3	4.93	6.88	2.82	2.11	4.77	2.79	2.26	3.18	2.20	2.88	2.60
Wt.% TiO_2	0.81	0.99	0.51	0.54	1.44	0.51	0.46	1.10	0.37	0.50	0.53
Wt.% FeO_T	3.17	4.13	2.64	2.66	5.17	3.34	2.21	2.90	3.81	3.42	3.38
Wt.% MgO	3.03	4.65	1.86	1.63	4.08	1.43	1.19	2.30	1.27	2.04	1.87
Wt.% CaO	2.39	5.56	1.32	1.19	3.07	1.80	1.89	1.14	1.54	1.60	1.47
Wt.% Na_2O	0.98	1.41	0.78	0.57	2.29	0.60	0.57	1.34	1.41	1.05	1.14
Wt.% K_2O	1.20	1.74	0.80	0.61	0.98	0.78	0.72	1.49	0.70	0.93	0.69
Ni (ppm)	508	590	416	444					255	386	424
Mn (ppm)	374	427	415	397				22,000[21]	351	503	557
Zn (ppm)	685	832	450	522				24,100[22]	227	407	372
Cr (ppm)	555	626	297	526					153	356	476
Co (ppm)	418	474	357	440					21	338	405

*The average number of components and standard deviation on components for each model across folds are given in supplementary document Table 4S.

**LANL models trained on a different, smaller, and less diverse training set of 408 samples with outliers removed for varying elements [8]

†Model described in Lanza et al. [21] using small data set, valid for compositions with <10 wt.% MnO

‡Model described in Lasue et al.[22] using small data set for predictions at 95% confidence level

FIGURES

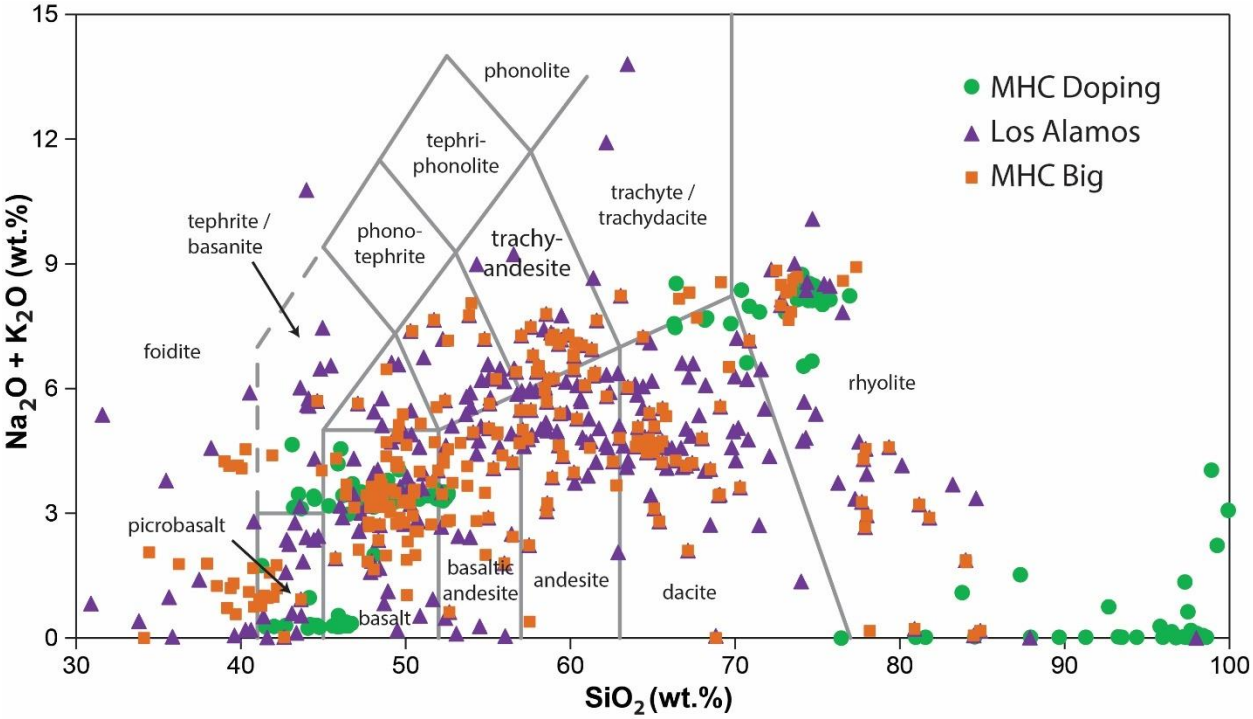

Fig. 1. Total alkali versus silica diagram showing the range of compositions in the three data sets studied here. Some analyses total to ~103 wt.% because of large error bars on SiO_2, K_2O, and Na_2O propagated onto the sums. Although this plot is conventionally used to show only volcanic rock compositions (for which names are given in regions shown here), it provides a convenient graphical representation of our range of compositions, which include both igneous and sedimentary rock types.

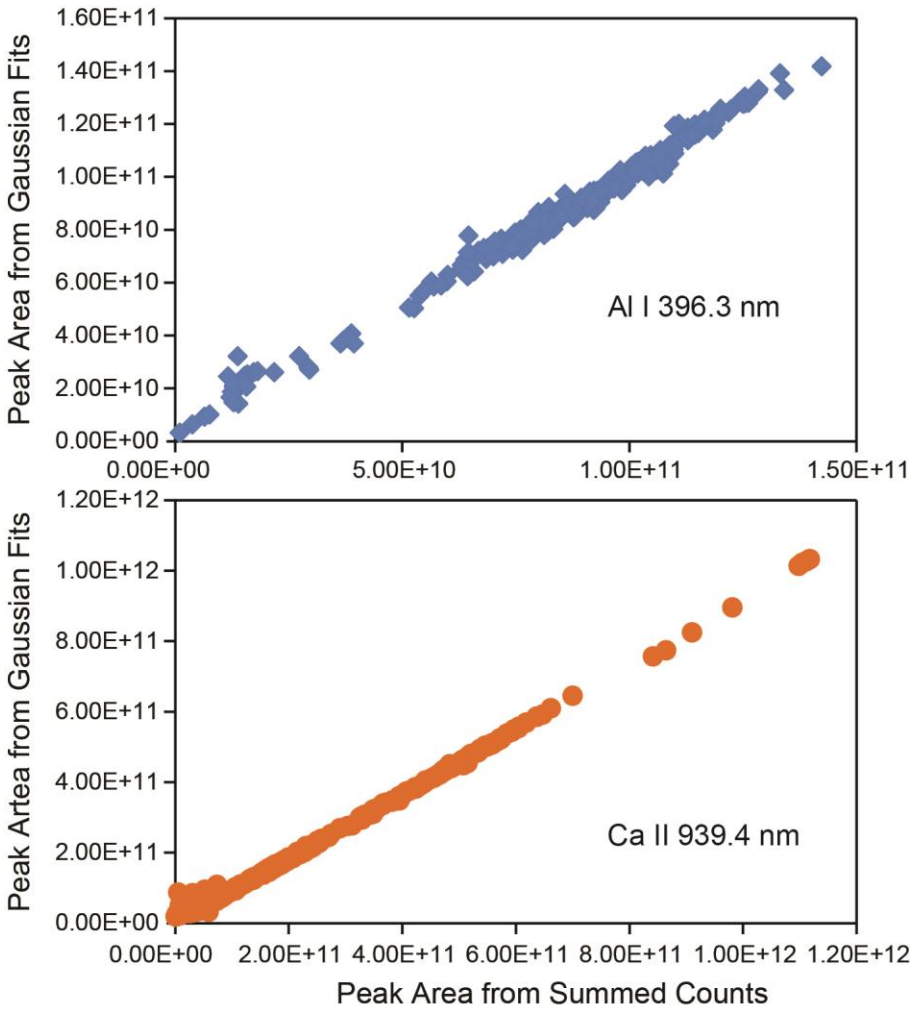

Fig. 2. Comparison of peak areas calculated using summed counts after custom baseline removal (*x* axis) against peak areas calculated by fitting the peaks using Gaussian peak shapes (*y* axis). The two methods yield comparable results, though there is some scatter at very small areas, which is to be expected from very small concentrations.

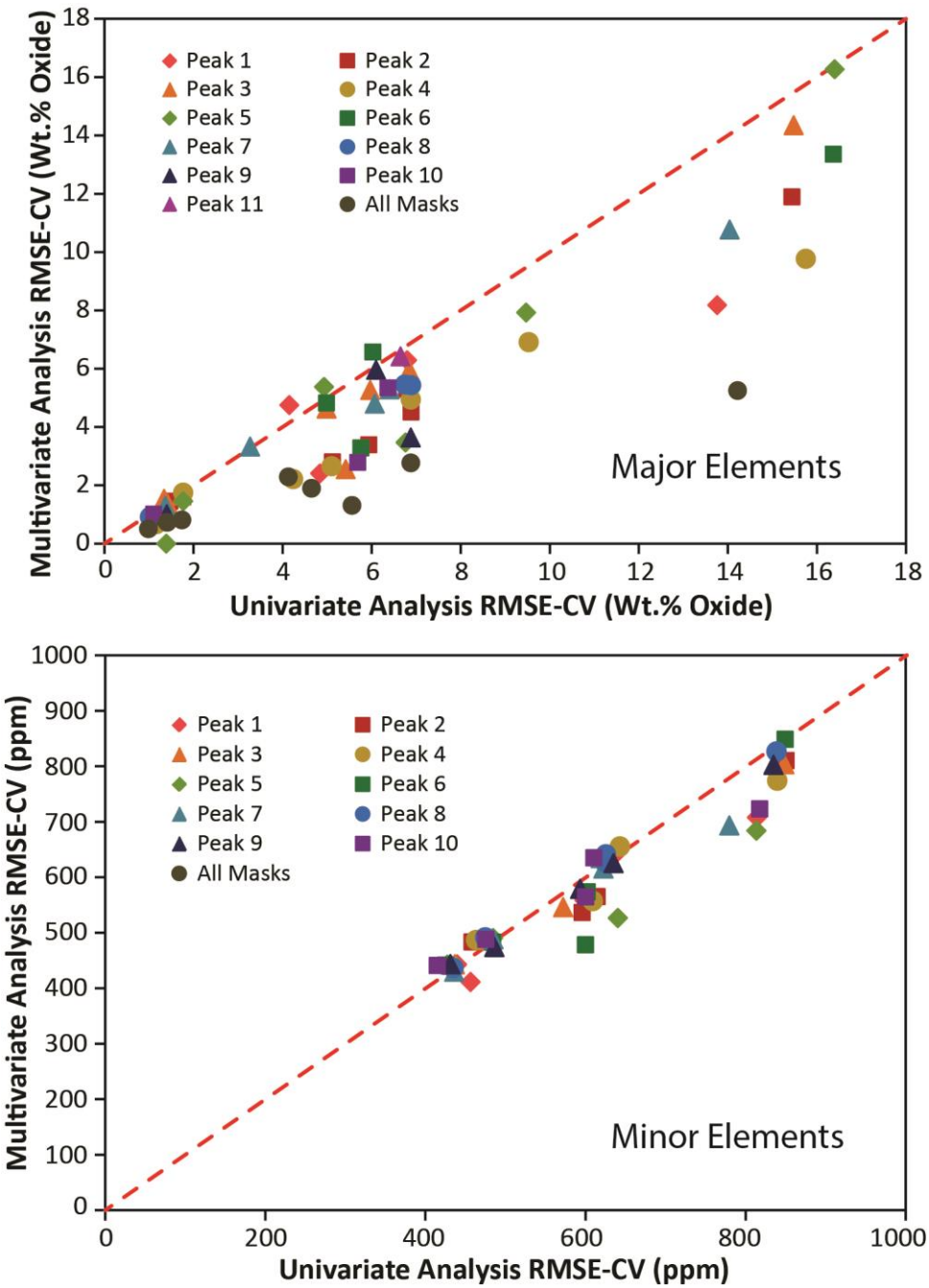

Fig. 3. RMSE-CV errors for univariate models compared with multivariate equivalents (PLS) for the same peak or multiple peaks. For major elements, multivariate models outperform univariate for nearly every case. For minor elements, univariate and multivariate models provide comparable results.

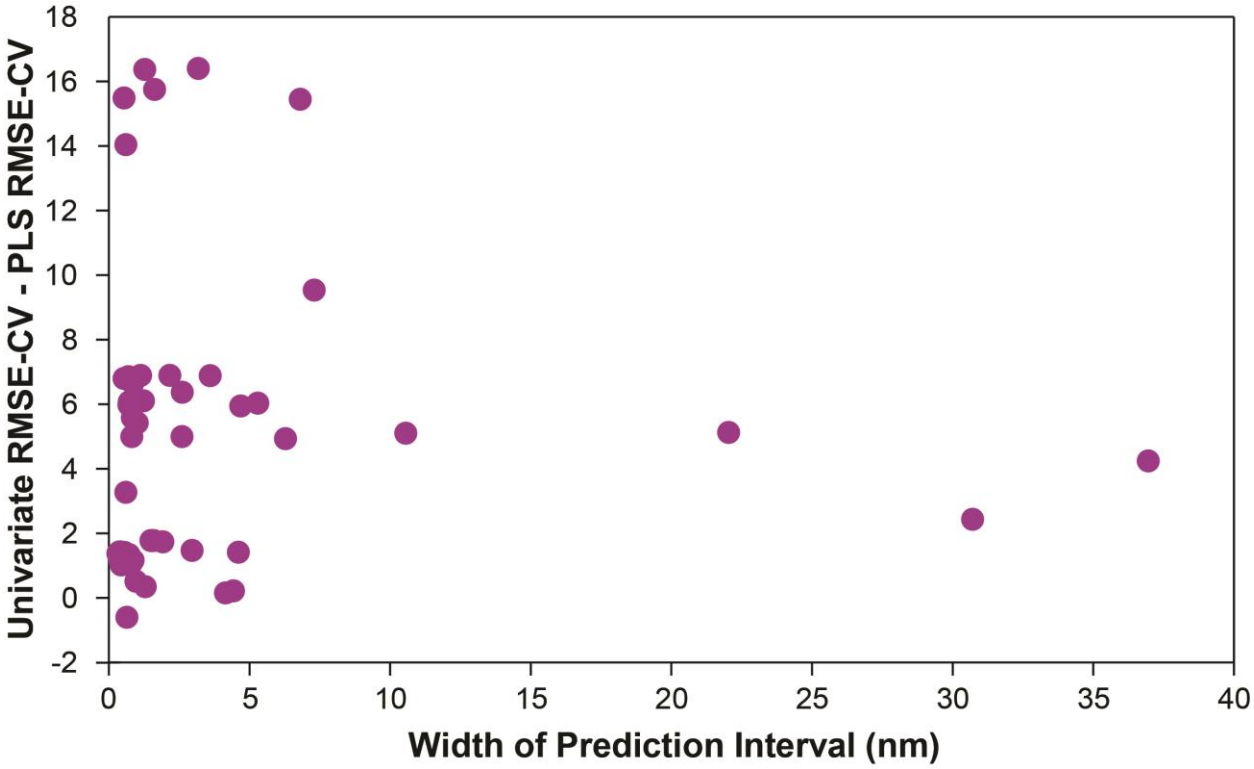

Fig. 4. Comparison of the number of channels (expressed in nm as the width of the interval used) used in univariate or multivariate predictions (the 57 ranges given in Tables 2 and 3) versus the improvement (reduced error bar) in accuracy for the two methods. The improvement of PLS accuracy over that of univariate is mostly independent of the width of the interval used.

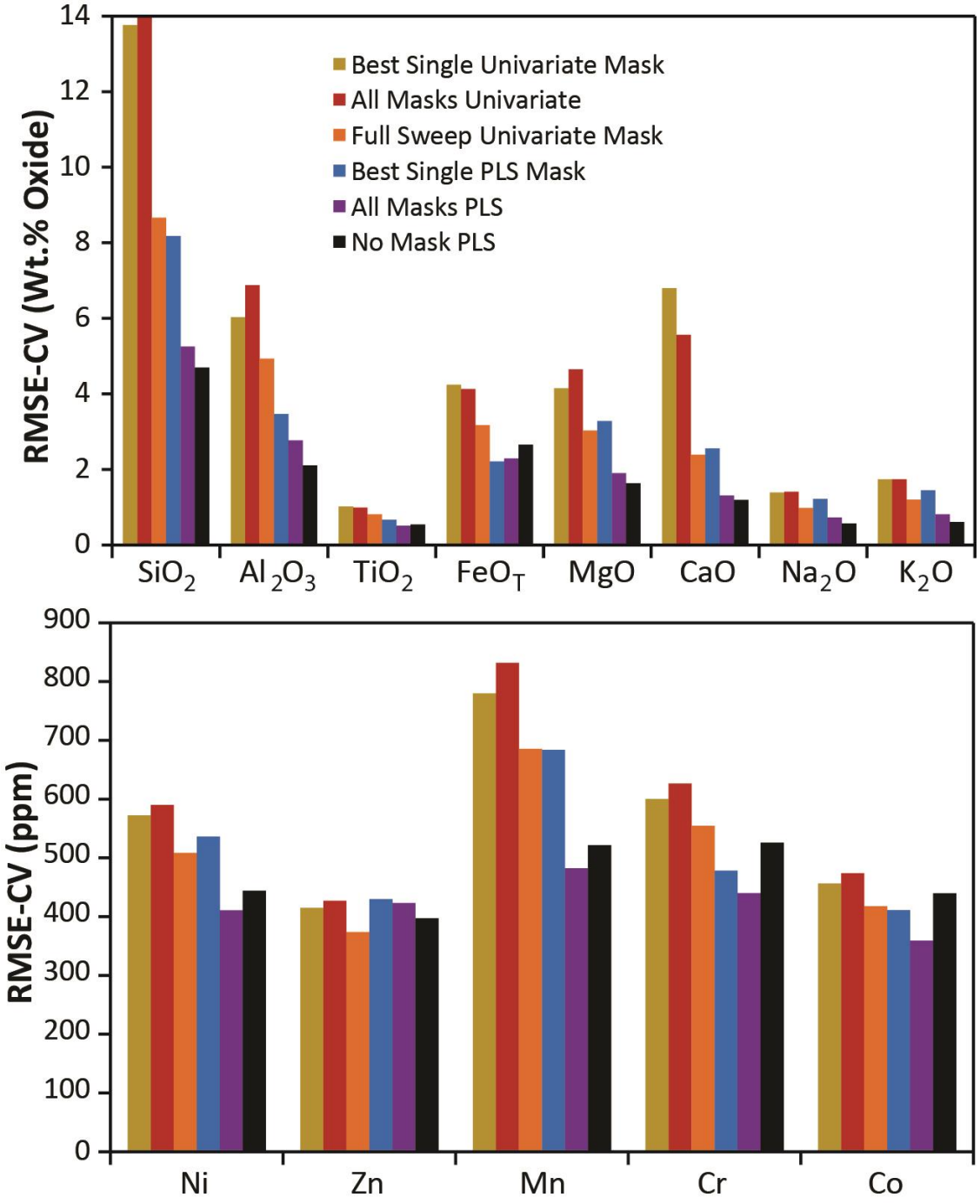

Fig. 5. Comparison of RMSE-CV values for six models using only data acquired at Mount Holyoke. Data from Tables 2 and 3. Best single models are the lowest RMSE-CV for any individual model. For major elements with many emission lines, masks are not needed. For minor elements, masking overcomes the effects of geochemical camouflage and focuses the predictions on regions where known lines from each element are found, generally resulting in more accurate predictions.

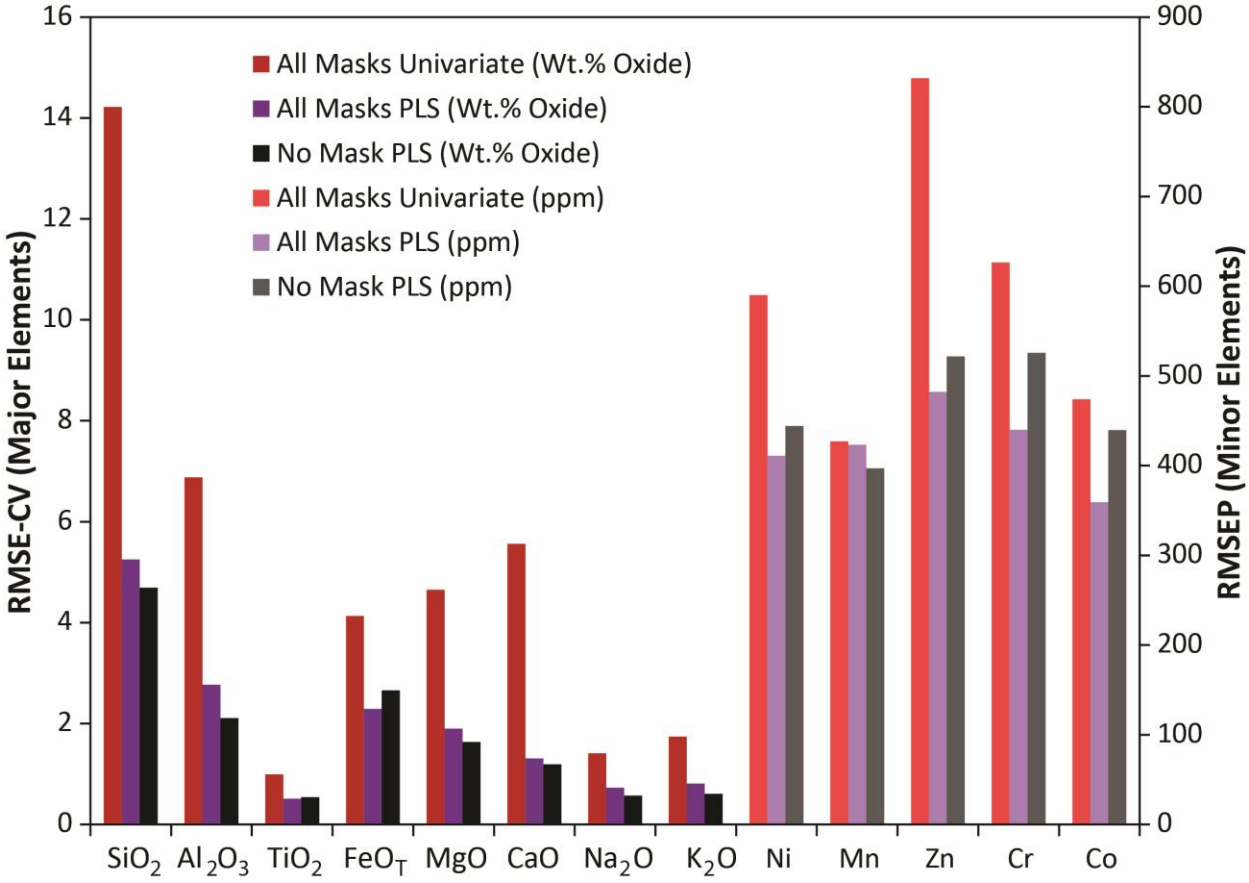

Fig. 6. Graphical illustration of the selected columns of Table 2 for data collected from two large datasets at multiple laser powers on the Mount Holyoke LIBS instrument. RMSE-CV values are plotted on the left axis for major elements and on the right axis for minor elements. For major elements, it is apparent that optimal accuracy is obtained when all spectral channels are included (no masking), likely because major elements have many peaks scattered throughout these channels and thus masking results in a loss of predictive information. For minor elements, masking gives comparable and sometimes better results than no masking and PLS consistently outperforms univariate analyses.

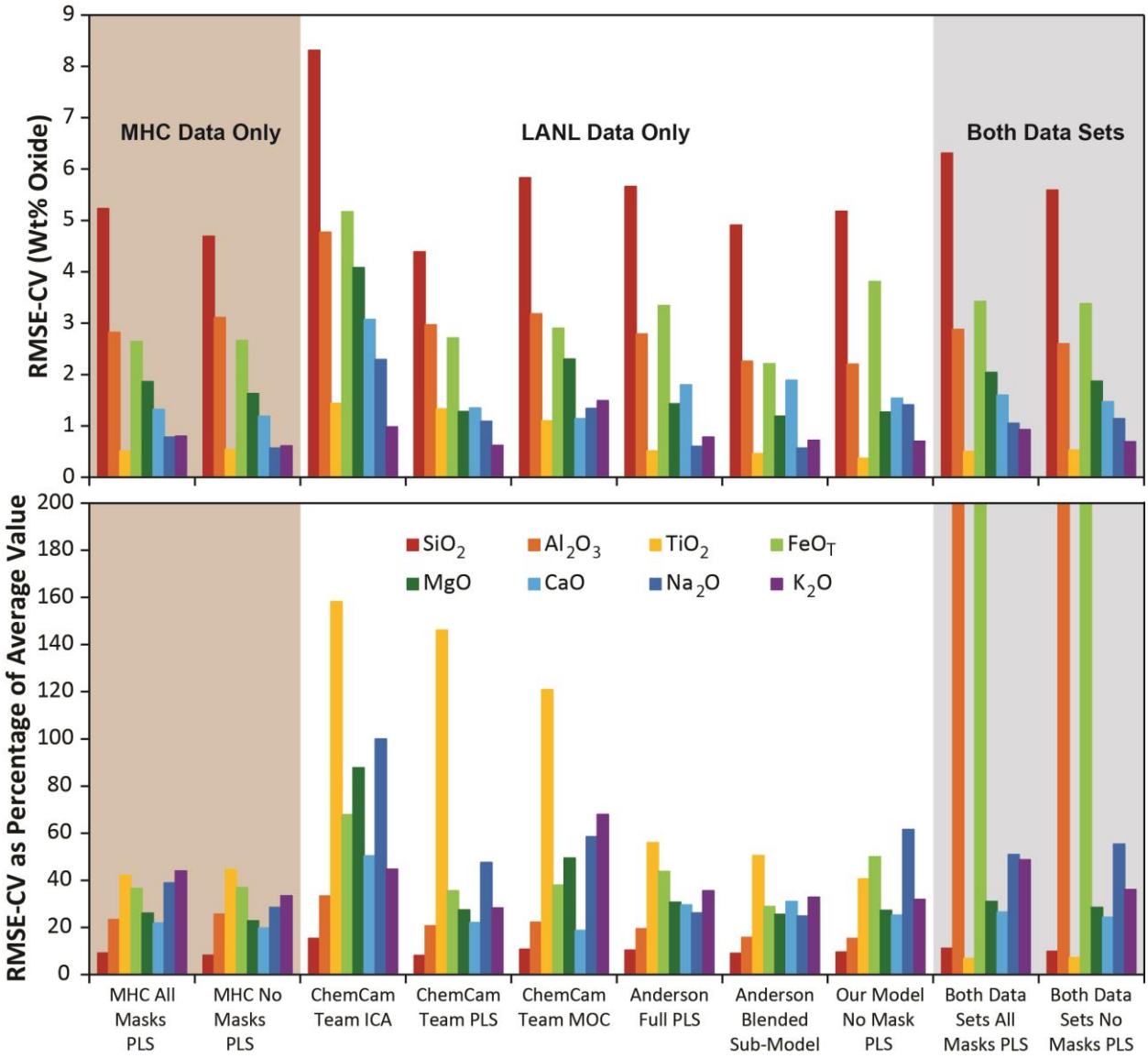

Fig. 7. Cross-validation root mean square prediction errors from PLS expressed as absolute values (top panel) and as percentages of the average value for each variable (bottom panel) for major elements comparing results from use of MHC data only, LANL data only, and the combined data set.

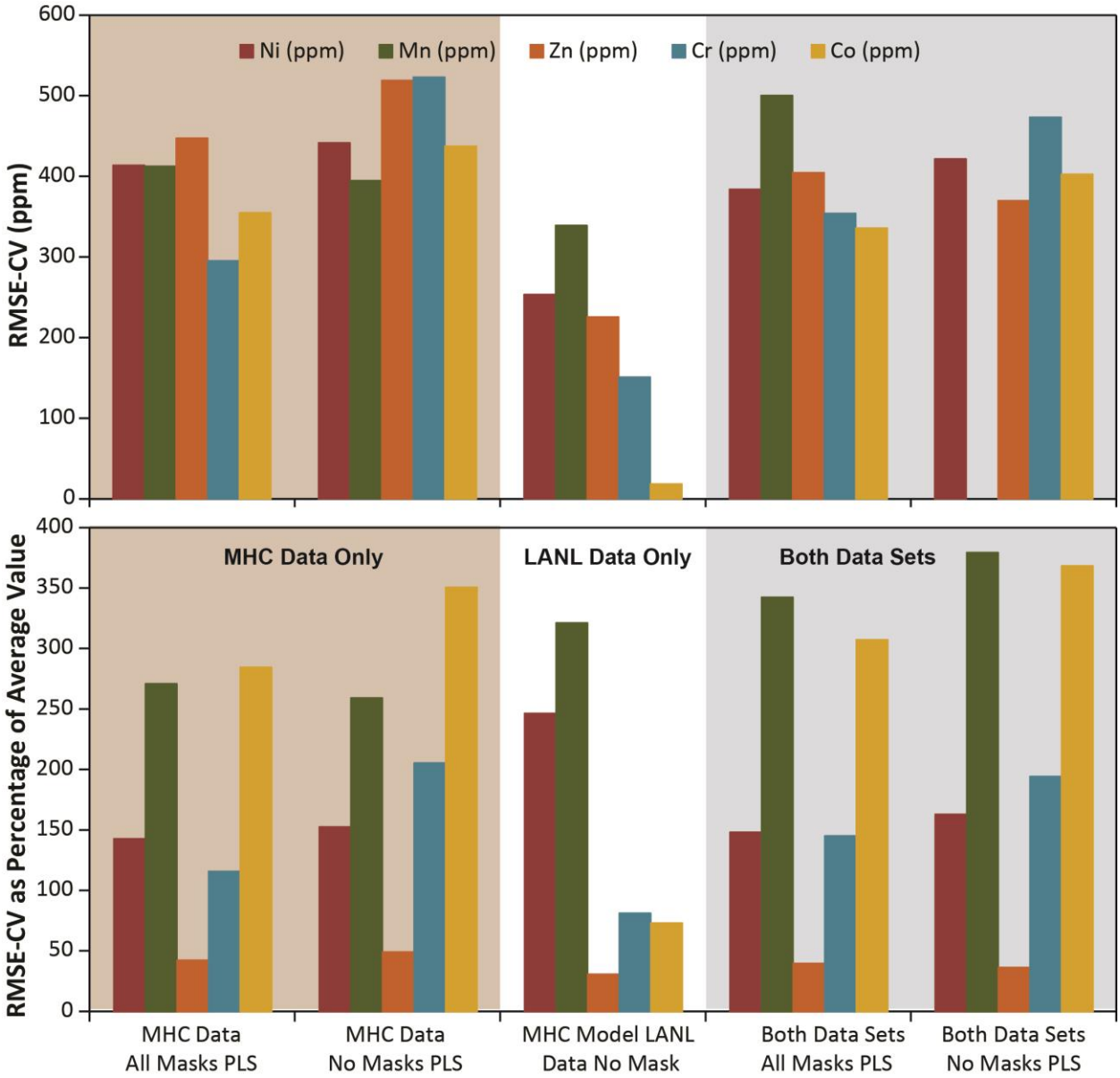

Fig. 8. Root mean square prediction errors expressed as absolute values (top panel) and as percentages of the average value for each variable (bottom panel) for minor elements comparing results from use of MHC data only, LANL data only, and the combined data set. The relative RMSE-CV values are quite large for these minor elements compared to the major elements shown in Fig. 6. Prediction errors do not change significantly when the LANL data are combined with the MHC data. Errors for Zn (shown on right axis of lower panel) are especially large because there are few useful Zn emission lines in the wavelength range of our spectrometers, As hypothesized by Lepore et al. [30], relatively large RMSE-CV values for the minor elements are likely limited by their inherently low signal to noise ratios, so that variations in prediction models have less of an effect on their accuracy.

Graphical Abstract