DEEP LEARNING MODELS FOR SPECTROSCOPIC DATA: SEMI-SUPERVIED GENERATIVE MODELS APPLIED TO LASER-INDUCED BREAKDOWN SPECTROSCOPIC DATA. Ian Gemp[1], Ishan Durugkar[1], Mario Parente[2], M. Darby Dyar[3], and Sridhar Mahadevan[1]. [1]College of information and Computer Sciences, University of Massachusetts, 140 Governor's Drive, Amherst, MA 01003, imgemp@cs.umass.edu, [2]Dept. of Electrical and Computer Engineering, University of Massachusetts, 151 Holdsworth Way, Amherst, MA 01003, [3]Mount Holyoke College, Dept. of Astronomy, 50 College St., South Hadley, MA 01075.

Introduction: Multivariate analysis techniques are widely used for extracting elemental compositions from LIBS spectra of geological samples [1,2]. Because the ratio of ground truth samples to the number of channels in each spectrum is small, the corresponding regression problem is underdetermined. Regularization (L_1, L_2) is essential for both better conditioning the problem and handling noise in spectral signals [3,4], but it ignores unlabeled data. A more robust approach embeds the data in a representation especially suited for the prediction problem. Boucher et al. [5] applied manifold learning for regression that embeds the samples using both sample and label similarity, outperforming previous manifold preprocessing techniques on LIBS data. But these approaches are limited by the expressive power of their models. In this project, we construct and train a deep generative model of LIBS data for the purposes of composition prediction and spectra generation. With this model, we generate spectra to validate our model assumptions. We conclude with a discussion of future work.

Choosing Deep Learning Models for LIBS applications: LIBS data have several important properties that must be considered in building a model. The physics of LIBS shots, including interactions between surface and sample and within the plasma itself, is complex and highly nonlinear. Compositions are continuous and spectral data are high-dimensional with typically >6000 channels. From publicly released ChemCam data, we have access to a large supply of unlabeled samples, as well as databases at Los Alamos National Lab [6] and Mount Holyoke College [7].

This application thus requires highly expressive, flexible semi-supervised generative models capable of capturing these nonlinear relationships. Deep generative models such as that described in Kingma et al. [8] fit this bill. However, we expand on that earlier work by considering not just unlabeled samples (those with unknown compositions, x) and sample-label pairs (x,y), but also labels (compositions) on their own (y). Our variational model allows efficient inference and learning through stochastic gradient descent as well as semi-supervised training of the whole Mars dataset. Normalizing flows [9] allow modeling of the complex posterior over the simplex (i.e., compositions domain). This approach allows end-to-end training of both regressor and generator (or decoder) through gradient descent making model and training adjustments relatively pain-free. The goal of these models is to extract compositional endmembers from LIBS data. Unlike many other types of spectroscopy used in planetary science where the unmixing task is to deconvolve contributions from different mineral phases, the end-members in LIBS are notional. Each represents the concentration of an element such as Si in the sample being analyzed. Once these endmembers have been identified, then quantification of their individual concentrations is enabled. In this project, we extract end-members for SiO_2, Al_2O_3, TiO_2, Fe_2O_3, MnO, MgO, CaO, Na_2O, and K_2O.

Model Description: Our model assumes the data can be explained by a generative process parameterized by θ: (compositions) $y \sim p(y)$, (nuisance) $z \sim p(z)$, (spectra) $x \sim p(x|y,z;\theta)$. Exact posterior inference is intractable, so we perform approximate inference with an approximate posterior, q_ϕ, and maximize a variational lower bound. Specifically, θ and ϕ parameterize deep feedforward neural networks that output the means and variances of isotropic gaussians representing p and q. See [8] and [10] for more details.

Results: The variational model described above was fit to the ChemCam LIBS data received from the Mars rover and publicly available on the Planetary Data Systems website. We obtained end-members by generating spectral samples (x) from corners of the simplex (y) with zero nuisance (z). **Figure 1** shows the spectral signal generated from setting the latent variable to 100% of selected variables of interest. Labels on those plots indicate which emission line corresponds to each element, and demonstrate that the model is picking the correct signal for each element. There is good agreement between the elemental signal generated by the model and the spectral signatures typically referenced in practice.

Future Work: This demonstration shows that our deep generative model successfully extracts emission lines from calibration target data acquired by ChemCam on Mars. The next step is to use these models for the purposes of composition prediction, and that research is ongoing.

Acknowledgments: We are grateful for support of this work from NSF grants NSF grants CHE-1306133, CHE-1307179, IIS-1564083, and IIS-1564032, and NASA grants NNX15AC82G and NNA14AB04A, the latter to the RIS[4]E SSERVI node.

References: [1] Clegg S. et al. (2009) *Spectroch. Acta B.*, *64*, 79-88. [2] Tucker J. M. et al. (2011) *Chem. Geol.*, *277*, 137-148. [3]. Kalivas J. (2012) *J. Chemom.*, *26*, 218-230. [4]

Chun H. and Keles S. (2010) *J. Royal Statist. Soc: B, 72*, 3-25. [5] Boucher T. et al. (2015) *Fifth AI in Space IJCAI Workshop*. [6] Clegg S. et al. (in press) *Spectroch. Acta B.* [7] Dyar M. D. et al. (2016) *LPS XLVII*, Abstract #2205. [8] Kingma D. P. et al. (2014) In *Advances in Neural Information Processing Systems*, pp. 3581–3589. [9] Rezende D. J. et al. (2015) In *Proc. of 32nd Intl. Conf. Machine Learning*, pp. 1530-1538. [10] Gemp I. et al. (2016) *arXiv preprint arXiv:1608.05983*.

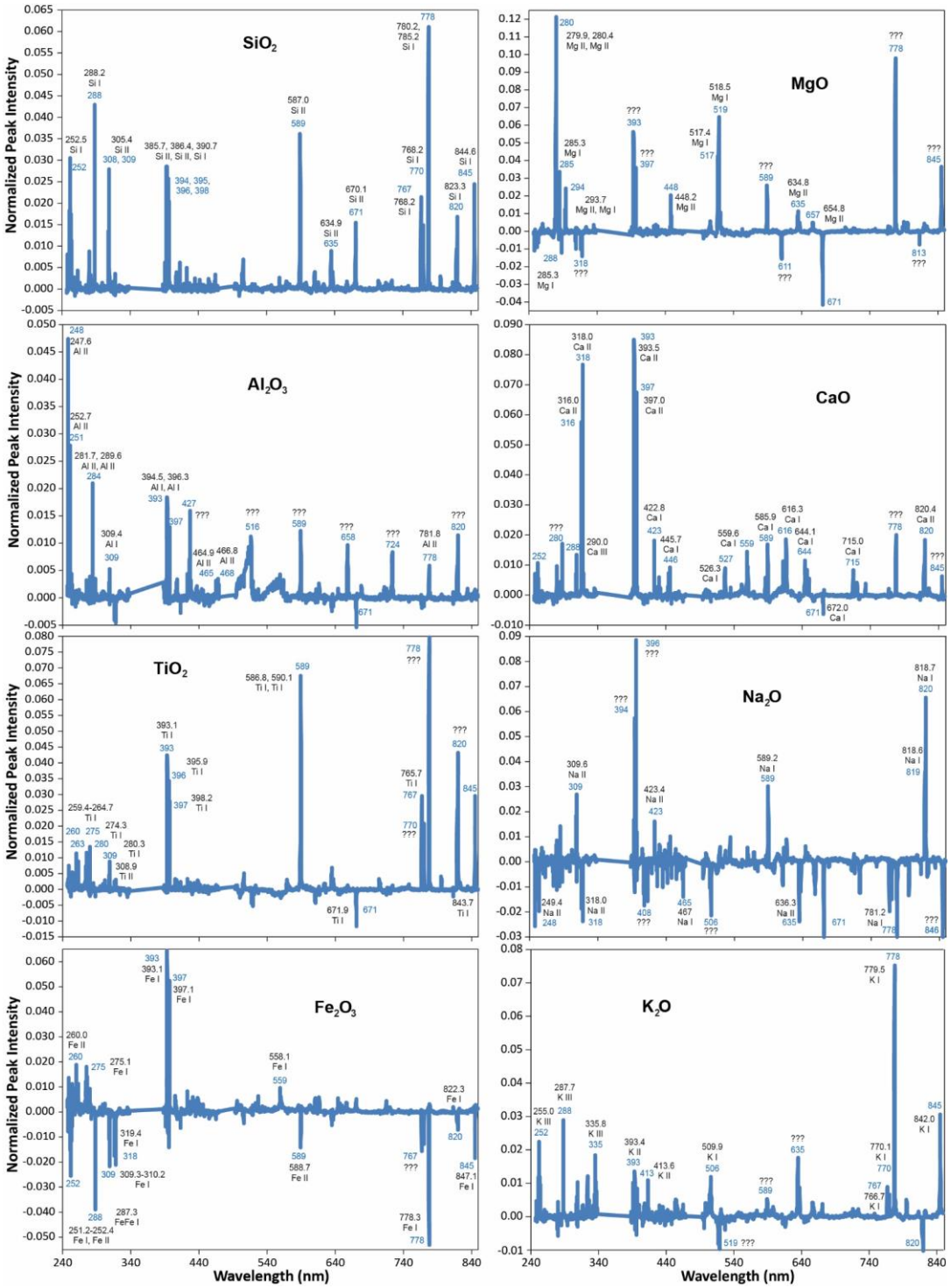

Figure 1. Spectral signals generated from our model conditioned on a composition of 100% pure oxides. Annotations in blue denote the locations of peaks in the generated signal while black denotes the location of known emission lines from each element. Such confirmations of the ability of the models to extract endmember components enable LIBS data prediction from compositions alone, as well as quantitative estimation of relative proportions of end-members.