MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Jessica Kong	Chemistry	University of Washington	kongjy@uw.edu

Project Title

Data-driven Analysis of Correlations between Chemical Structure and Electrical Function at the Nanoscale

Project Synopsis (approx. 100 words)

The goal of this project is to accelerate our understanding of how chemical composition and nanoscale structure are related to properties like electrical conductivity and work function that ultimately give rise to materials with promising applications in energy harvesting and storage. This project will also establish nanoscale chemical mapping as an integral part of materials characterization. We aim to develop software that will enable pixel-by-pixel analysis of multimodal images and as a result, enable the discovery of relationships between local chemical composition and electronic functions in materials on the nanoscale. We will examine synthetic image data, then use model polymer blends, and finish with new halide perovskite semiconductors.

Identified Data-Science Collaborative Need (approx. 100 words)

I am seeking data scientists with backgrounds in dimensionality reduction, hyperspectral unmixing, and multivariate statistical learning techniques. We will utilize tools that provide physically meaningful results (like nonnegative independent component analysis and nonnegative matrix factorization) to determine the spectral components and fractional abundances of materials in hyperspectral image data. To analyze these results, we will use a Bayesian approach to both linear and nonlinear multivariate regression techniques such as multiple linear regression and multivariate adaptive regression splines. *The project is also open to other techniques that team members believe will be insightful in understanding the relationship between chemical structure and electrical properties.* Ultimately, we will incorporate these techniques into Pycroscopy, an existing open source, Python package currently under development by nanoscale imaging communities.

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

I prepared all samples and collected all images on an Asylum MFP3D Atomic Force Microscope (conductivity and potential maps) and Molecular Vista VistaScope (hyperspectral infrared maps). I have translators for proprietary files from both instruments that can be utilized to read the data into Python as a class. Both raw files and translators can be shared with the data science teams.

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

The central goal of this project is to develop an open source, distributable software package that will enable pixel-by-pixel analysis of multimodal images and subsequent discovery of correlations between local chemical composition and electronic functions in materials on the nanoscale. A recently developed scanning probe technique called photoinduced force microscopy has helped make nanoscale chemical maps easy to acquire.[1] This project utilizes this advancement to accelerate our understanding of how chemical composition and structure are related to electrical properties that ultimately give rise to materials with promising applications in materials from solar cells to batteries. We will incorporate the software we create with Pycroscopy, an existing software package for image processing and scientific analysis currently under development by communities using nanoscale imaging techniques.[2]

To examine the relationships between chemical structure and electrical properties, we will look at three atomic force microscopy (AFM) data sets of varying complexities as shown in Figure 1. The synthetic data set contains simulated hyperspectral and functional images, the latter with a known relationship to the hyperspectral image. Both the polymer blend of poly(methyl methacrylate) (PMMA) and poly(3-hexylthiophene) (P3HT) and

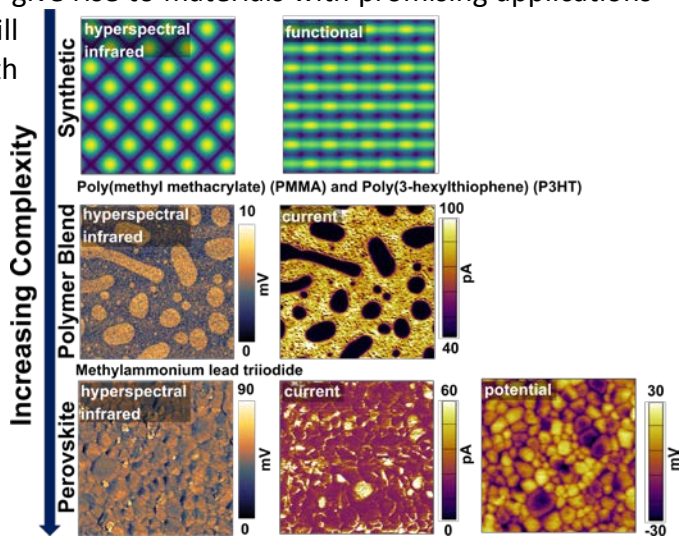

Figure 1. Data sets of varying complexities to be used for discovering relationships between chemical structure and electrical properties on the nanoscale.

methylammonium lead triiodide perovskite system contain a hyperspectral (~36 million spectroscopic points) and at least one electrically functional (conductivity and/or potential) image (~65 thousand points, each). The scope of this project can be divided into three main sections:

1. Image Registration
2. Dimensionality Reduction and Hyperspectral Unmixing
3. Linear and Nonlinear Multivariate Regression

each of which will include incorporating these functionalities into Pycroscopy with Python.

Image Registration

This aspect of the project will focus on writing wrapper functions in Python that will integrate capability to register images with Pycroscopy.

In AFM-based techniques, a nanometer-scale probe is raster scanned over a micron-sized region while simultaneously measuring topography and a local property (e.g, electrical current or chemical signature) with nanometer scale resolution. Due to differences in tip orientation and torsion, piezo hysteresis, and instrument drift, even consecutive images from the same instrument on ostensibly the same physical location need to be aligned. Dipy[3], an available Python package, can be used to implement affine transformations on the topography image to sufficiently align the functional maps as shown in Figure 2 for the polymer blend system. This capability will be incorporated with Pycroscopy, which does not yet have this functionality.

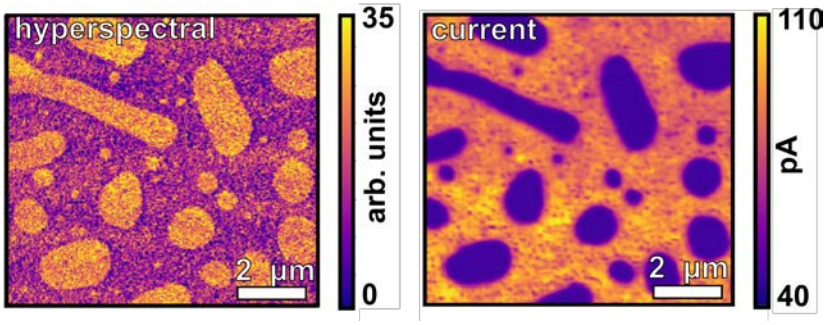

Figure 2. Hyperspectral and conductivity images aligned using Dipy's affine transformation capabilities.

Dimensionality Reduction and Hyperspectral Unmixing

This part of the project will assess the applicability of nonnegative dimensionality reduction techniques and develop them for use on hyperspectral infrared images within Pycroscopy.

The hyperspectral images contain sub-diffraction limited infrared (IR) spectra, which can be thought of as a chemical fingerprint, at every pixel in a micron-scale image. The dimensionality of these images are 256 x 256 pixels with 559 spectral dimensions; each spectral dimension contains an intensity value associated with a wavenumber between 800 and 1800 cm^{-1}. Since every molecule has distinct vibrational modes that manifest as peaks at particular wavenumbers, IR spectra can be used to characterize the composition of a material.

With hyperspectral infrared images of a material, it can be particularly insightful if we can determine the corresponding spectra of constituent materials, their fractional abundances and their relationship with correlated electrical properties. Principal Component Analysis (PCA) can be helpful in distinguishing insulating PMMA aggregates from the semiconducting P3HT matrix

in a blend as shown in Figure 3(a) and (b). Spectral signatures must be positive to align with physical reality but by the nature of PCA, the principal components (PCs) are not always meaningful endmember spectra (Figure 3 (c) and (d)). This constraint makes nonnegative dimensionality reduction techniques such as nonnegative matrix factorization (NMF)[4], nonnegative independent component analysis (ICA)[5], and others of particular interest. While these techniques are suitable for simpler systems like the polymer blend, for materials with considerably more intermixed morphology and therefore more complex structure-function relationships, more novel techniques such as minimum volume constrained nonnegative matrix factorization[6], used for highly mixed image data may be required. ICA, NMF, and various matrix factorization techniques can, in principle, be implemented using packages like scikit-learn,[7] Python Matrix Factorization (PyMF)[8], and several others. Many of these, however, have been in the alpha development stage for several years and even those that are current, are limited in capability.

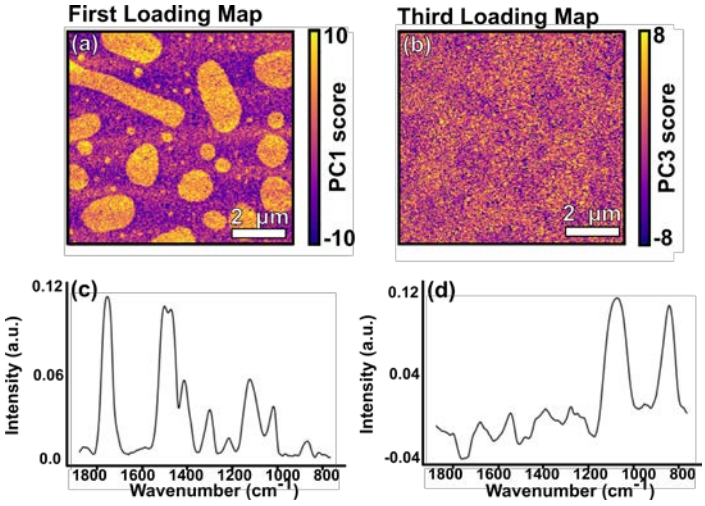

Figure 3. (a) and **(b)** Principal Component Analysis loadings for the **(c)** first and **(d)** third principal components, respectively.

Linear and Nonlinear Multivariate Regression

We will apply a Bayesian approach to linear and nonlinear multivariate regression techniques including multiple linear regression and multiple adaptive regression splines.

More importantly, we seek to understand the relationship between local chemical composition (the spectra of constituent materials and their fractional abundances) and function (a quantitative measure of a property). Prior work with multiple linear regression (MLR) where the first five principal components are regressed onto the current gives a model from which we can reasonably predict the current as shown in Figure 4. While qualitatively valid, the prediction from this initial approach does not capture intra-aggregate conductive domains (bright regions within dark aggregates in Figure 4a). To improve this prediction, we will turn to more general techniques such as multiple adaptive regression splines (MARS) which can flexibly model more sophisticated relationships by including nonlinearity and interaction terms.[9] Incorporating prior information about the systems can also improve the models we

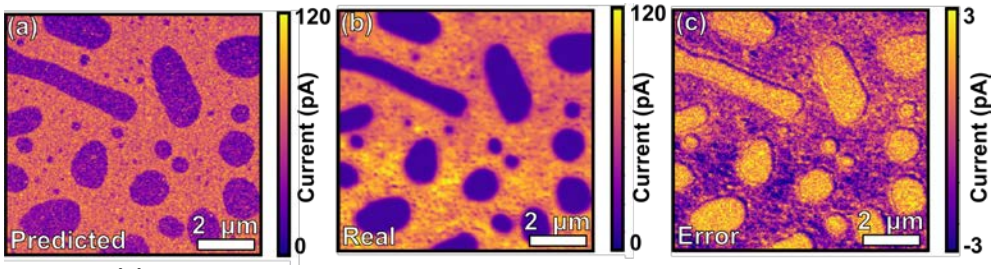

Figure 4. (a) Current as predicted with a multiple linear regression model obtained by regressing the first five principal components onto the current. **(b)** Real current as measured using conductive AFM. **(c)** Error image obtained by subtracting the real from predicted current.

develop; therefore, a Bayesian approach to these multivariate techniques may be more appropriate.[10] We can utilize and build on existing packages to implement MLR (scikit-learn[7]), MARS (py-earth[11], Orange[12]), and their Bayesian solution (PyMC[13], emcee[14], PyStan[15]).

References

(1) Nowak, D.; Morrison, W.; Wickramasinghe, H. K.; Jahng, J.; Potma, E.; Wan, L.; Ruiz, R.; Albrecht, T. R.; Schmidt, K.; Frommer, J.; Sanders, D. P.; Park, S. *Sci. Adv.* **2016**, *2* (3), e1501571.

(2) Somnath, S.; Jesse, S.; Laanait, N. *https://github.com/pycroscopy*.

(3) Garyfallidis, E.; Brett, M.; Amirbekian, B.; Rokem, A.; Van Der Walt, S.; Descoteaux, M.; Nimmo-Smith, I. *Front. Neuroinformatics* **2014**, *8*, 1–17.

(4) Paatero, P.; Tapper, U. *Environmetrics* **1994**, *5* (2), 111–126.

(5) Plumbley, M. D. *IEEE Trans. Neural Netw.* **2003**, *14* (3), 534–543.

(6) Miao, L.; Qi, H. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45* (3), 765–777.

(7) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(8) Thurau, C. *pymf: Python Matrix Factorization Module*; 2017.

(9) Friedman, J. H. *Ann. Stat.* **1991**, *19* (1), 1–67.

(10) Christian P. Robert. In *The Bayesian Choice*; Springer Texts in Statistics; Springer, New York, NY, 2007; pp 507–530.

(11) *py-earth: A Python implementation of Jerome Friedman's Multivariate Adaptive Regression Splines*; scikit-learn-contrib, 2018.

(12) Orange – Data Mining Fruitful & Fun https://orange.biolab.si/.

(13) Patil, A.; Huard, D.; Fonnesbeck, C. J. *J. Stat. Softw.* **2010**, *35* (4), 1–81.

(14) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. *Publ. Astron. Soc. Pac.* **2013**, *125* (925), 306–312.

(15) PyStan http://mc-stan.org/users/interfaces/pystan.