2025 ML/Microscopy Project Report

Fatima Anwar, Hayden Dennison, Julie Schlanz, Emily Stump
University of Cincinnati, Department of Chemistry

**All presented work was completed during the 2025 Microscopy Hackathon*

Identifying the Spatial Domains of Sm-doped $BiFeO_3$ using STEM and Machine Learning

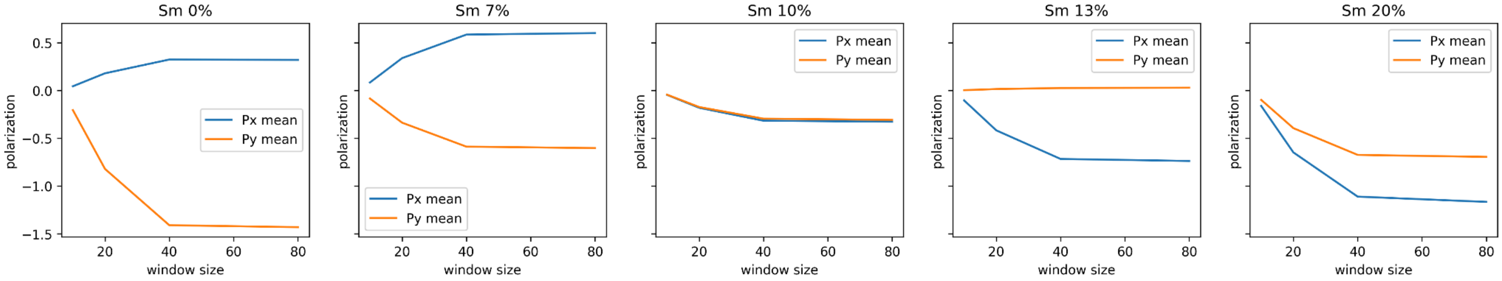
Background

Ferroelectric materials are defined by a permanent electric polarization. Unless specifically prepared otherwise, samples will exhibit multiple spatial domains that differ in the direction of polarization. With the use of scanning transmission electron microscopy (STEM) these domains can be identified. For the perovskite $BiFeO_3$,doing so involves locating each unit cell and observing its central iron atom displacement. This is a laborious process that is subject to error from multiple sources. For this reason, efforts have been made to use machine learning (ML) models (specifically deep models such as neural networks) to identify these domains purely from the STEM image data.[1]

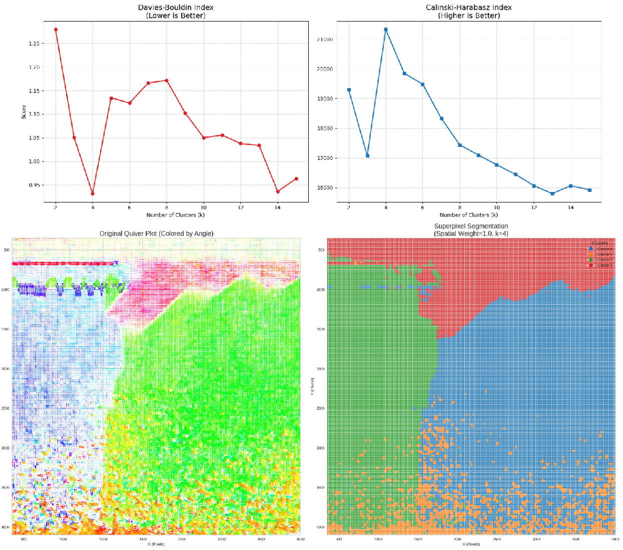
A previous study used principal component analysis (PCA) to visually reveal the domain walls of $BiFeO_3$.[2] Our study also uses PCA in addition to K-means clustering and shallow tree-based models to create an automated and more efficient method. It also investigates varying the window size and concentrations of samarium (Sm) doped $BiFeO_3$. We utilized a dataset made available in the Hackathon's Github repository, which consisted of STEM image data for a sample of $BiFeO_3$ with increasing levels of doping by Sm. Labels for the polarization domains were only provided for a small piece of the data for 0% Sm. Doping disrupts the local structure of the rhombohedral $BiFeO_3$ crystal and induces a shift to an orthorhombic crystal structure.

Methods/Results

The large image was first normalized and then broken up into smaller sub-images (patches) by sliding a square window of a defined size. Window sizes of 5, 10, 20, 40, and 80 pixels were done separately, each with a stride of half the window size (or equal to the window size when 5 was used). Using polarization vector components for each unit cell, average polarization components for each sub-image were calculated. The average polarization components across all sub-images was plotted for each Sm concentration (0%, 7%, 10%, 13%, 20%) against window size, where it was found that this value converged for window size 40 or larger.



For each window size, the patches for all Sm concentrations were combined into a single dataset and PCA was completed. It was found that the variance explained by the first 2 PCs converged at window size 40, much like the average polarization. When visualizing the patches in PC1xPC2 space there is some separation between Sm concentrations only for window size 40 and below. As mentioned above, only a small portion of the data was explicitly labelled in the notebook. Without labels for the Sm-doped images, it is difficult to evaluate the performance of any supervised learning method. With polarization vectors provided, this can be done manually but is time-intensive. For this reason, we explored the usage of K-Means clustering to identify domain labels automatically. We concatenated polarization components for each patch using their position. Multiple values of k were tested, and the Davies-Bouldin and Calinski-Harabasz indices were calculated to select the cluster number. Plots for 0% Sm and window size 40 are shown below, along with a comparison between the resulting clustering and the quiver plot of polarization vectors (vectors colored by angle with positive X-axis). We see good agreement, confirming this method. For smaller window sizes the clustering is not successful when compared with the quiver plot. For this reason, we chose to assign cluster labels to the small windows based on the label of the 40-size window that overlaps them.



Random forests with 100 trees were trained on each window size for 0% Sm to classify into labels. 5-Fold CV was done and the average test performance across all folds was evaluated. It was found that the results were similar regardless of window size, with accuracies of ~75%. Models had the greatest difficulty with the orange cluster along the bottom, which we interpret as an edge effect. All methods utilized the sklearn Python library for implementation[3].

Discussion

Our work demonstrates two major points. One, the ideal window size among those tested was found to be 40. Although the supervised methods performed well, the unsupervised methods struggled to distinguish domains and Sm concentrations for larger and smaller windows. This may be due to the lack of convergence for polarization on smaller windows and larger windows having long-range details that are unimportant for this task. Second, our work implies that random forests have some capability in this task even though they use flattened inputs that don't respect the geometry of pixels within a patch.

References

[1] Nelson, Christopher T., et al. "Deep learning ferroelectric polarization distributions from STEM data via with and without atom finding." *npj Computational Materials* 7.1 (2021): 149.

[2] (1) Borisevich, AlbinaY.; Ovchinnikov, O. S.; Chang, H. J.; Oxley, M. P.; Yu, P.; Seidel, J.; Eliseev, E. A.; Morozovska, A. N.; Ramesh, R.; Pennycook, S. J.; Kalinin, S. V. Mapping Octahedral Tilts and Polarization Across a Domain Wall in BiFeO3 from Z-Contrast Scanning Transmission Electron Microscopy Image Atomic Column Shape Analysis. ACS Nano 2010, 4 (10), 6071–6079. https://doi.org/10.1021/nn1011539.

[3] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.