Electronic Supplementary Material (ESI) for Digital Discovery.

Supplementary Information

Accelerating Nano-XANES Imaging via Feature Selection

Samantha Tetef[1], Ajith Pattammattel[2], Yong S. Chu[2], Maria K. Y. Chan[3], Gerald T. Seidler[1]

[1] University of Washington, Seattle, WA, 98195, USA

[2] National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY, 11973, USA

[3] Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois, 60439, USA

Table of Contents

Figure S1 RFE results on linear combinations of experimental data

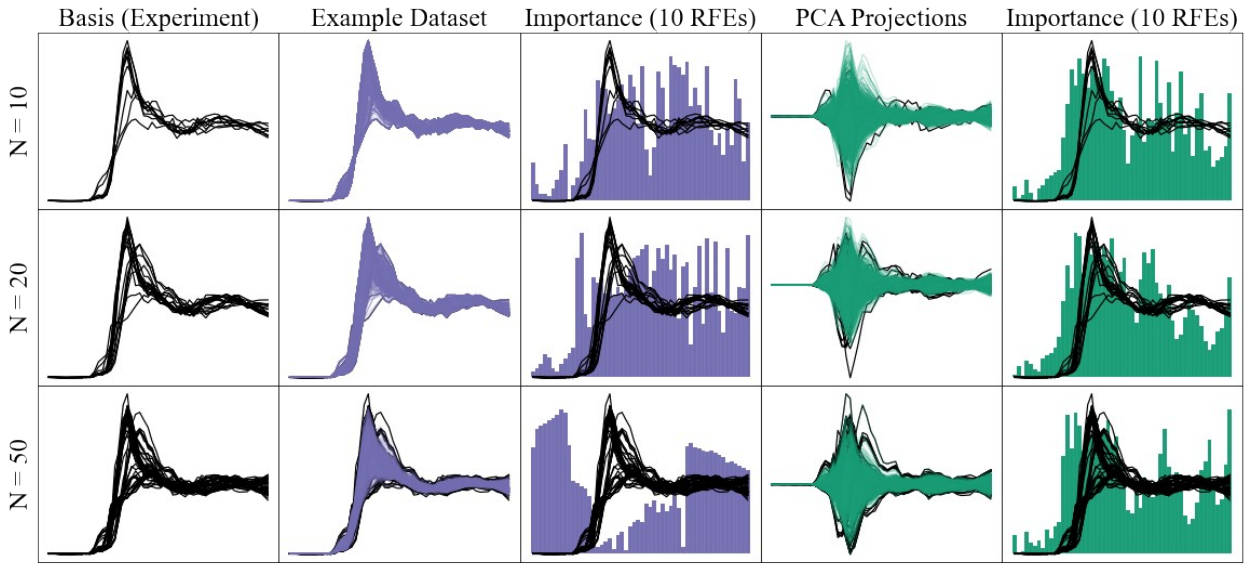

Figure S1 (a) Random sampling of experimental data to act as a basis for linear combinations of spectra. Note that there is no guarantee that the basis spectra span or equally sample the experimental domain. (b) 1000 linear combinations generated from the corresponding basis. (c) The compiled results of an ensemble of 10 RFEs trained on the spectra. (d) The equivalent dataset except projected onto the first six principal components. (e) The compiled results of an ensemble of 10 RFEs trained on the principal components. When N (the basis set size) is 50, there is so much linear dependence in the basis set that the RFE fails because it chooses points in the pre-edge, which has no variation.

Figure S2 RFE results on gaussian basis sets

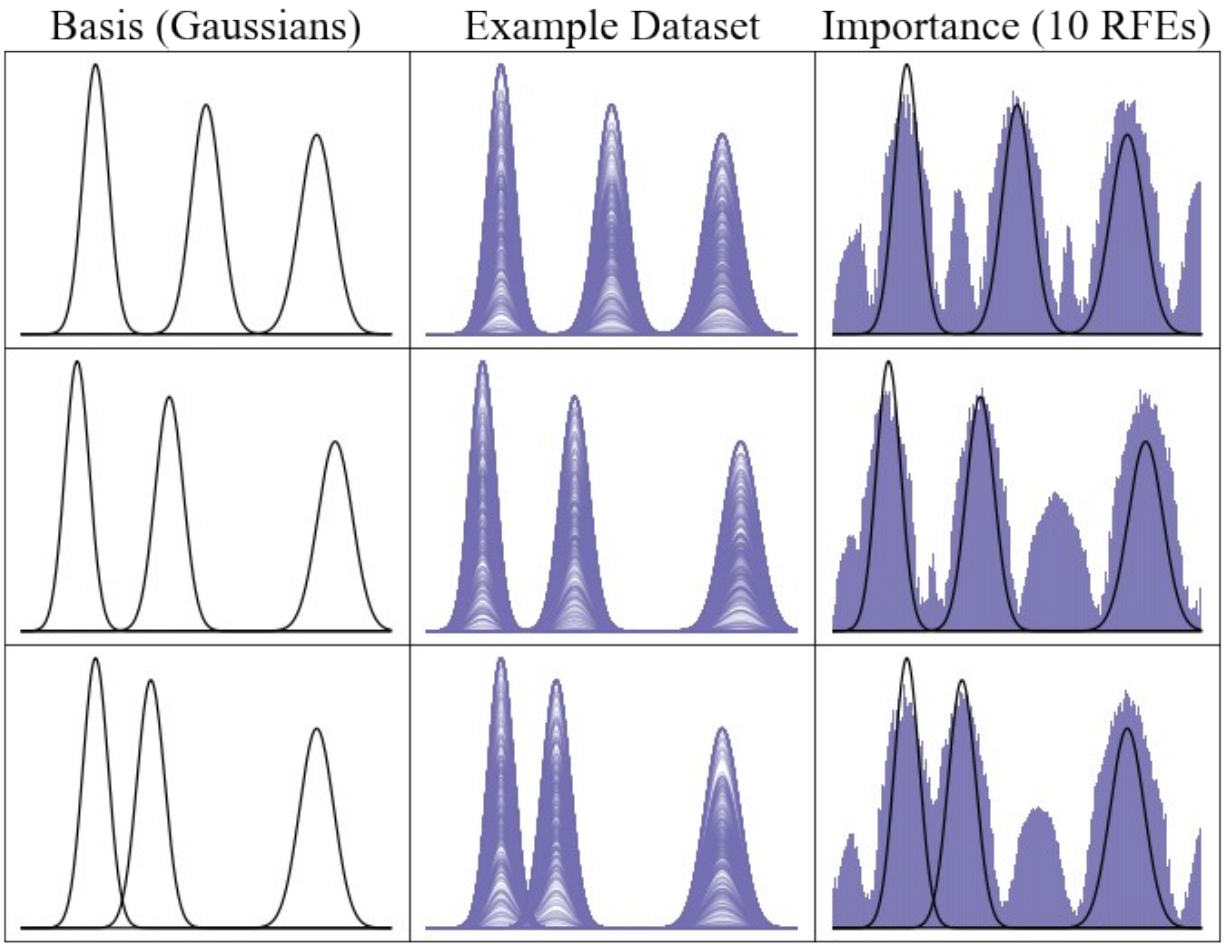

Figure S2 Test of the RFE by using three gaussians used as basis spectra (left-most column) to make linear combinations (middle column). The RFE clearly picks features (i.e., "energies") that correspond to the highest variation. We used linear regression as our base estimator. However, after the regions corresponding to the three distributions are filled, the RFE must rank areas in between peaks where there is no signal. The peaks in importance between the Gaussians represent random selections in these regimes.

Figure S3 RFE results on both linear and nonlinear input/output pairs

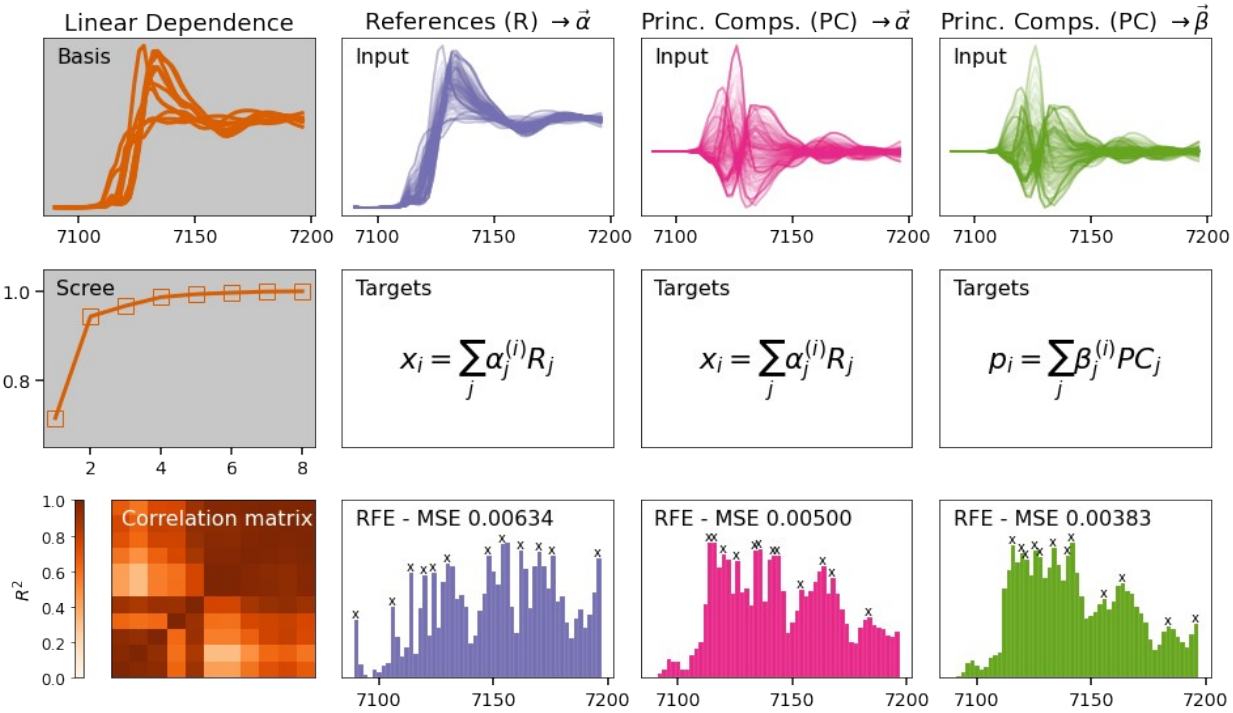

Figure S3 Comparing results of linear and nonlinear inputs to the RFE. The scree plot and correlation matrix both demonstrate the linear dependence of the reference spectra. The RFE results, where linear combinations of spectra are the input and the concentrations that created those spectra are output, is shown in purple. Instead, using projections onto the first few principal components as input (with the same output) is shown in pink. The green shows the RFE results using both linear input and outputs. The total results for an ensemble of 10 RFE algorithms for each are shown at the bottom, along with the mean squared error (MSE) of predictions using LASSO linear combination fitting (LASSO-LCF) on a generated dataset of linear combinations of reference spectra.

Figure S4 Scree plot showing PCs needed for increasing reference library size

Figure S4 The number of principal components (PCs) needed to explain 99% variance of the reference set. Starting with the four known references, we randomly selected additional references from the set of 11 total references used in this study. After the 11 references were chosen, we randomly selected additional references from another larger set of 64 Fe K edge XANES to constitute the reference library. We reselected these random additions 50 times and show the average and standard deviation of the calculated number of principal components for that reference library size. This large Fe K-edge XANES reference library was taken from M. Marcus and P. Lam, *Environmental Science* **2014**, *11* (1), 10-17.

Figure S5 Correlation matrices of references

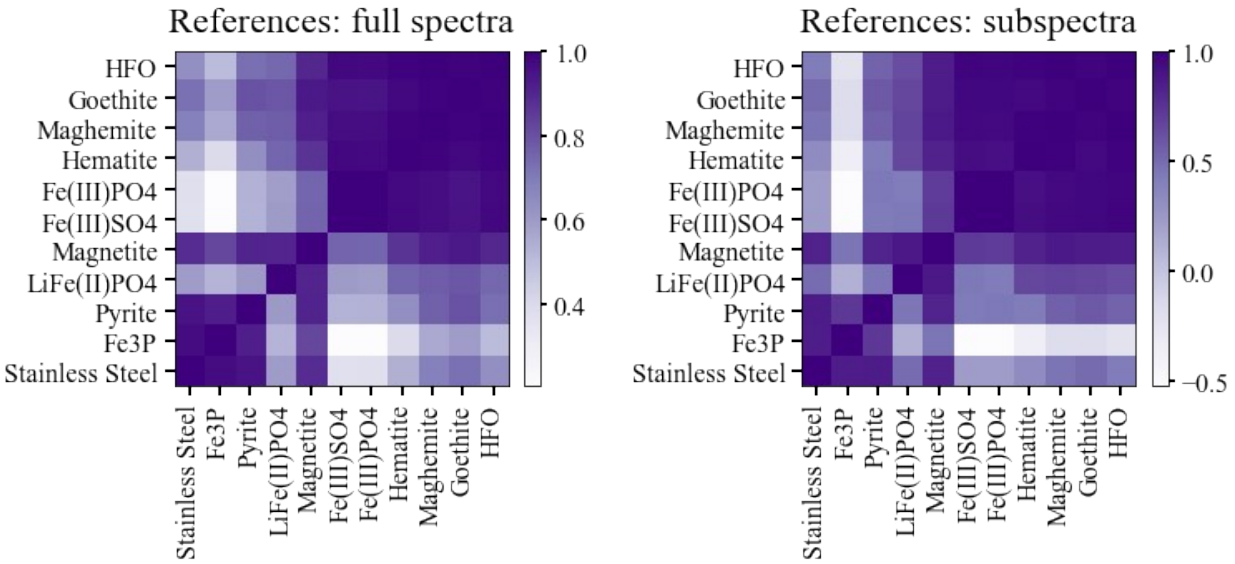

Figure S5 Correlation, or similarity matrices, of the reference set for both the entire spectra and the 14-energy subspectra. The correlation coefficient (R^2) qualitatively looks the same for both, although the quantitative range for the subspectra is larger, indicating global correlations (and information) is retained.

Figure S6 Scree plot of experimental data

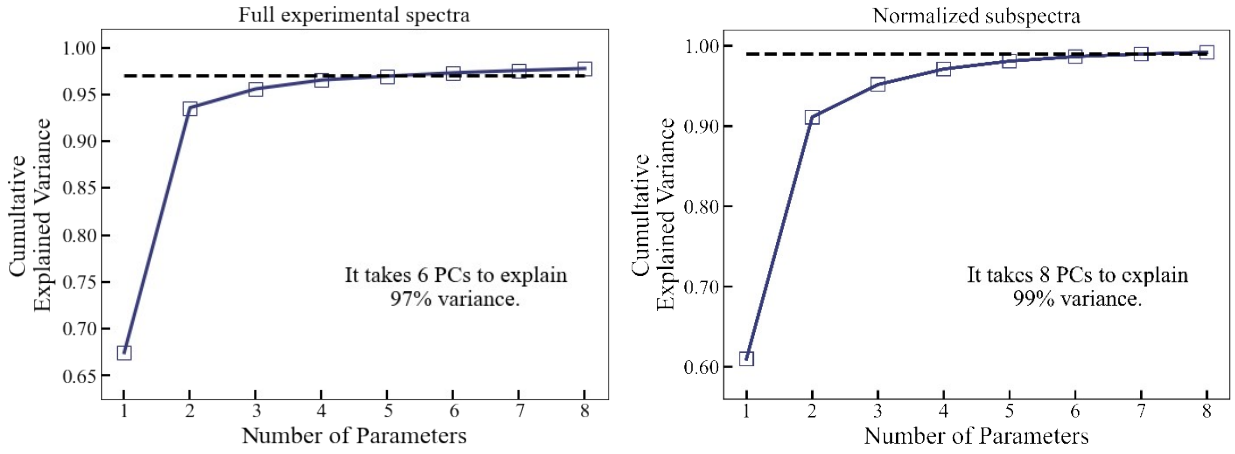

Figure S6 Scree plot of experimental data on full spectra (top) versus subspectra (bottom).

Figure S7 First four PCs of subspectra

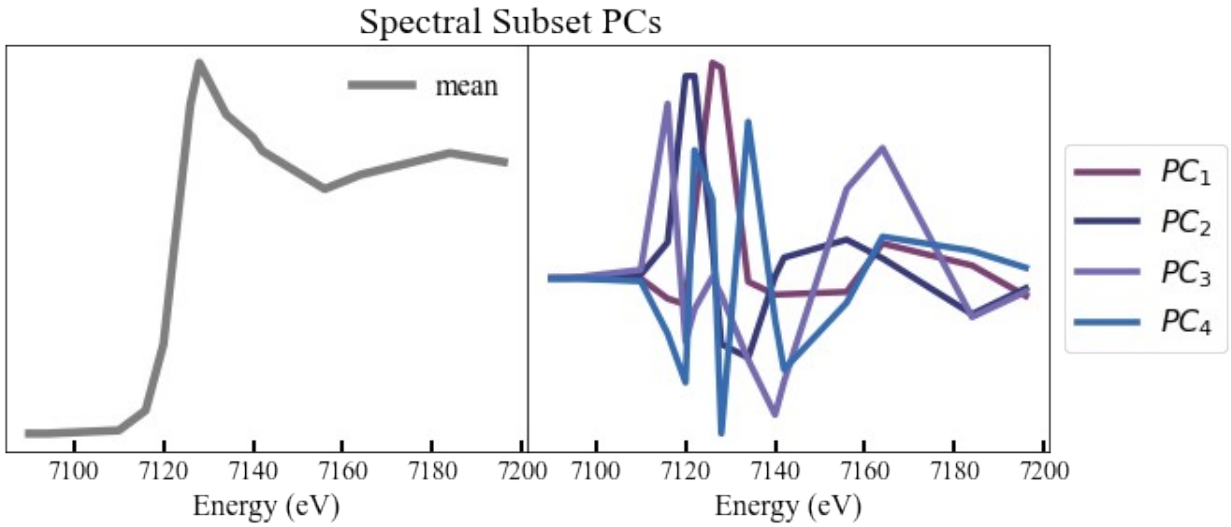

Figure S7 First four principal components of the 16-energy point subspectra. These components, in theory, should match with the principal components from the full spectral dataset, if all information is retained.

Figure S8 PCA triangle plot on subspectra

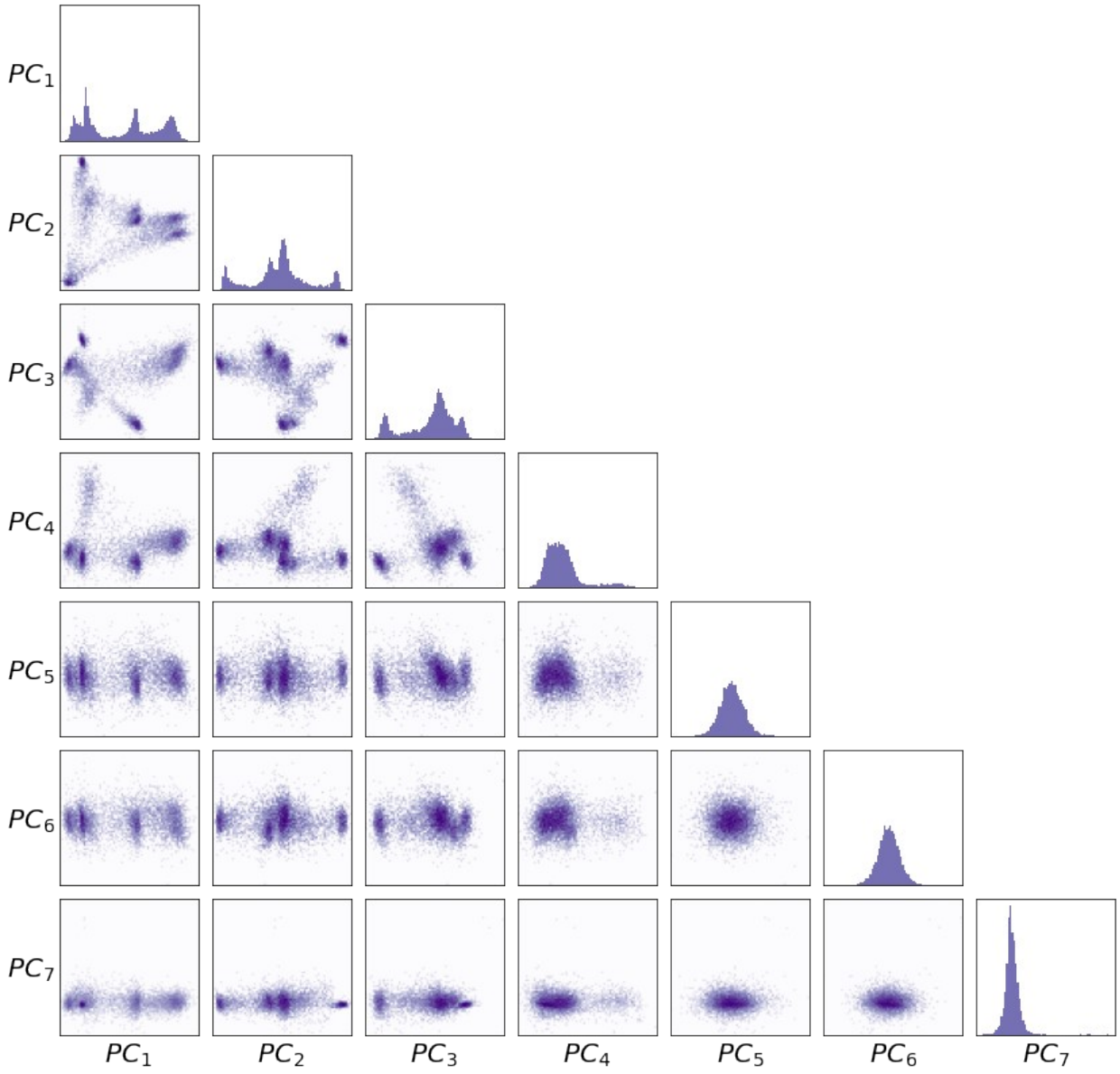

Figure S8 PCA triangle plot of experimental data on subspectra.

Figure S9 UMAP and dbscan clustering on subspectra

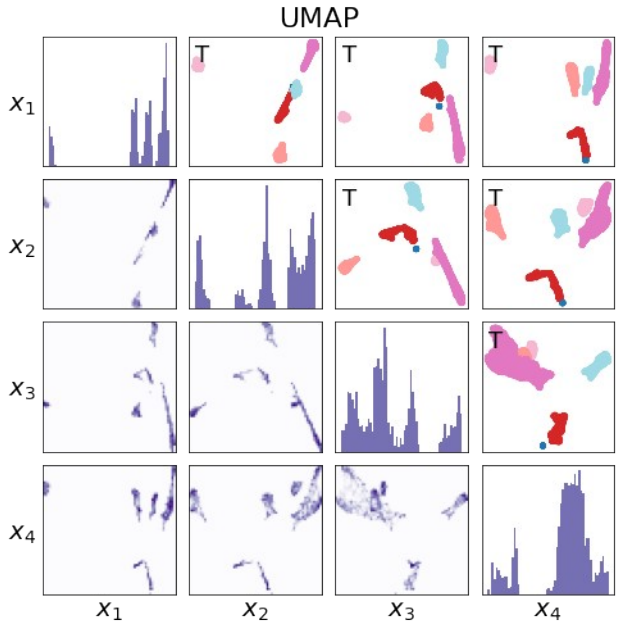

Figure S9 UMAP and dbscan on energy subset.

Figure S10 Effects of varying hyperparameters that control spatial grouping

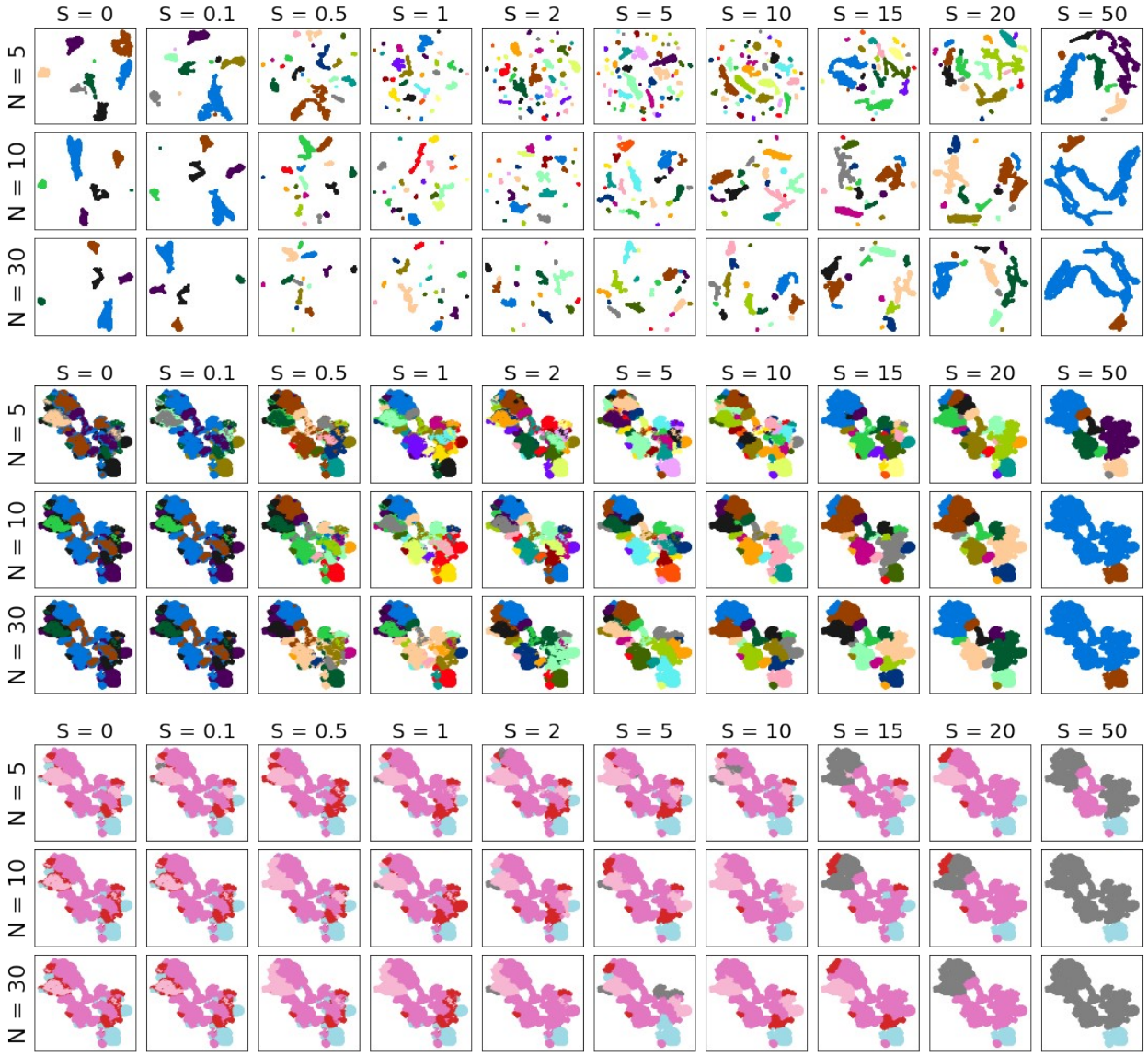

Figure S10 Strength of spatial encoding (S) versus UMAP's number of neighbors (N). The minimum distance in UMAP is 0 and dbscan epsilon is 1 for all. The top section shows the UMAP space color-coded by dbscan clusters, the middle shows the same clusters but on the 2D map, and the bottom shows the max contributions from the LCF fits. Pink = Pyrite, magenta = LFP, blue = Hematite, red = SS, and gray = all other references.