Manifold Projection Image Segmentation for Nano-XANES Imaging

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

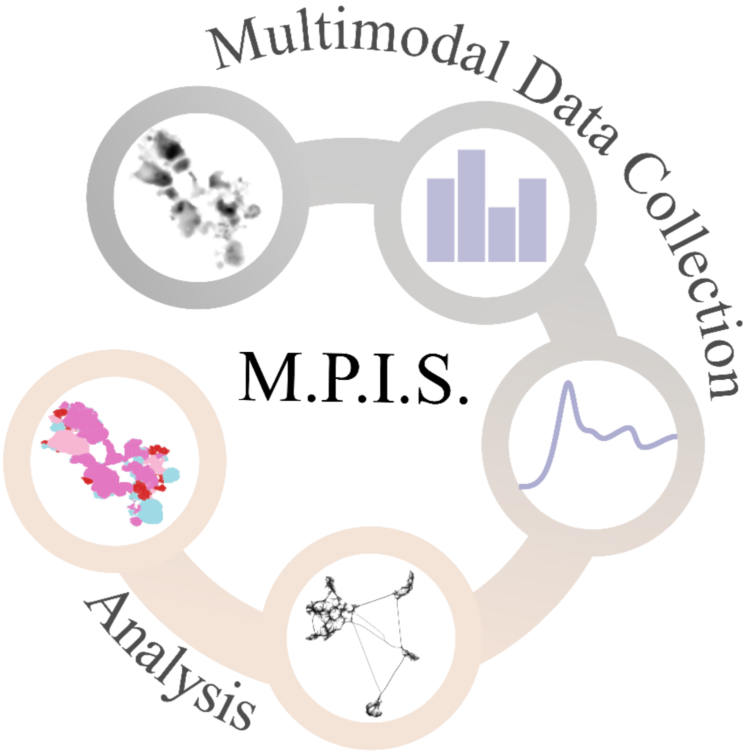
*mchan@anl.gov, *seidler@uw.edu

Abstract

As spectral imaging techniques are becoming more prominent in science, advanced image segmentation algorithms are required to identify appropriate domains in these images. We present a version of image segmentation called manifold projection image segmentation (MPIS) that is generally applicable to a broad range of systems without the need for training because MPIS uses unsupervised machine learning with a few physically motivated hyperparameters. We apply MPIS to nano-XANES imaging, where X-ray Absorption Near Edge Structure (XANES) spectra are collected with nanometer spatial resolution. We show the superiority of manifold projection over linear transformations, such as the commonly used Principal Component Analysis (PCA). Moreover, MPIS maintains accuracy while reducing computation time and sensitivity to noise compared to the standard nano-XANES imaging analysis procedure. Finally, we demonstrate how multimodal information, such as X-ray Fluorescence (XRF) data and spatial location of pixels, can be incorporated into the MPIS framework. We propose that MPIS is adaptable for any spectral

imaging technique, including Scanning Transmission X-ray Microscopy (STXM), where the length scale of domains is larger than the resolution of the experiment.

TOC Graphic



1. Introduction

The increased popularity in various scientific fields of utilizing high-throughput imaging techniques, especially spectral imaging experiments, has benefited from advanced image segmentation algorithms so that researchers can identify regions in the image belonging to the same domain, object, phase, etc. Image segmentation methods that utilize multimodal characterization measurements as input, which potentially could be high-dimensional, are especially beneficial for the scientific community[1-4]. However, most common image segmentation algorithms utilize either hand-crafted rules or convolutional neural networks, both of which can suffer from lack of generalizability. Moreover, not enough training data or unreliable data simulations may make inference unreliable when using neural networks. An alternative is to utilize manifold projection and clustering based on spectral similarity rather than deep learning, effectively performing *semantic* image segmentation. This manifold projection image segmentation, which we will refer to as MPIS, has seen success when applied to mass spectroscopy images[5] and flow cytometry data[6].

Here, we apply MPIS to a hyperspectral imaging technique called nanoscale X-ray absorption near edge structure (nano-XANES)[7-12]. XANES is a common experimental technique in materials science, chemistry, and biology as it is sensitive to local electronic structure around a chosen atomic species[13]. The goal of nano-XANES imaging is often to generate a compositional map of the local coordination, or phase, of the element of interest. The most common practice to make these maps is to perform linear combination fitting (LCF) to a reference library of spectra at every pixel, treating the image as an ensemble of independent XANES spectra and ignoring the spatial location of each spectrum. The analysis of XANES spectra using LCF is highly constrained by the prior knowledge of the system as well as the limited information encoded in the spectra. Uncertainty in the system can lead to an overly large library with poor linear independence. Only after LCF are the fit results used to construct the spatial phase maps. This approach is slow by requiring many fits, and any errors in the LCF fitting process are propagated when creating the spatial maps.

We show that MPIS has three major benefits compared to the above standard practice. First, by implementing MPIS, we flip the order of generating spatial domains and identifying the compositions of those domains. By switching this order and decoupling the image segmentation from the LCF, one can substitute improved or specialized classification or regression techniques as needed while maintaining persistent image segmentation, or domain identification, via MPIS. Therefore, phase maps are independent of the selection of a reference library and any errors in the LCF results. Second, MPIS can cluster the reference library in the context of the experimental spectra. Because the reference library can be a large set of numerically similar spectra, researchers often report LCF fits by grouping reference spectra by chemical class. MPIS instead provides a data-driven way to group references together.

Furthermore, MPIS is adaptable for encoding multimodal data into the image segmentation pipeline. In almost all nano-XANES imaging studies, XRF maps are simultaneously acquired for every XANES spectrum, producing a higher dimensional dataset enabling both spectroscopic and elemental analysis. We show that MPIS applied to an augmented encoding of both XANES and XRF spectra can better separate low signal-to-noise from high signal-to-noise data. Furthermore, encoding the position of the pixel into MPIS can generate smaller domain regions – divided by spatial location rather than global groupings – that is more akin to instance image segmentation, for example separating out each physical particle in the same phase. Finally, to force sparsity in the fits, the standard LCF practice uses stepwise regression, i.e., performing regression on all enumerated subsets of the reference library. We instead substitute stepwise regression with

LASSO regression, as presented in Jahrman et al.[14], to speed up computations. Finally, we perform LCF on the cluster-averaged spectra specified by MPIS. By having a data-informed way to average spectra together without losing spatial resolution, our LCF is more robust against noise.

We propose MPIS can be broadly applied to a wide range of spectroscopy techniques, including multimodal experiments and other imaging techniques such as Scanning Transmission X-ray Microscopy (STXM). While we demonstrate MPIS on a nano-XANES image, MPIS can be used to cluster any ensemble-based measurement because the pixel location in the image is encoded as optional multimodal information. Furthermore, MPIS decreases the chances of overfitting by requiring fewer, and physically meaningful, hyperparameters compared to deep learning. Finally, MPIS increases the reliability and efficiency of high-throughput analysis by speeding up computations and reducing sensitivity to noise in subsequent analysis.

2. Methods

2.1 Experimental Methods

Our sample is composed of Lithium iron phosphate (LFP), pyrite, hematite, and stainless steel. The Lithium iron phosphate (LFP) and pyrite (Pyr) samples were purchased from Sigma Aldrich, St Louis, MO. Hematite (Hem) and stainless-steel (SS) nanoparticles were obtained from US Nano Research (Houston, TX, USA). A heterogeneous mixture of the above-mentioned particles was created by physically mixing in acetone, followed by ultrasonication for 5 minutes. About 5 mL of the dispersed mixture was drop-casted onto a silicon nitride membrane (Norcia, Edmonton, Canada) and the solvent was dried in air. All data was collected at the Hard X-ray Nanoprobe (HXN) Beamline at National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory[15, 16]. A detailed methodology for nano-XANES acquisition was previously published.[12]

We measured Fe K-edge XANES using nano-XANES imaging, where our data consist of a 3D image with 155 x 160 spatial pixels and just over 70 photon energies between 7.08 to 7.20 keV, or approximately 25,000 spectra. Further processing of the stack was performed via the XMIDAS program[17]. The energy stack was first spatially aligned using the image registration tool in XMIDAS that uses the PyStackReg package[18]. Spectra are preprocessed via normalization and alignment using the standard procedures as described in Ref. [19]. Finally, the spectra were assembled to create a 3D array (Fe energy stack) for XANES analysis. For each scan point, an energy X-ray Fluorescence spectrum was collected with a three-element silicon drift detector (Vortex, Hitachi Inc) positioned at 90 deg to the sample. The XRF spectra were processed using the PyXRF software[20] to compute elemental maps.

2.2 Computational Methods

Uniform Manifold Approximation and Projection (UMAP)[21] was implemented using the umap-learn Python package. UMAP requires two main hyperparameters – the number of neighbors (to control cluster sizes and thus global versus local similarity) and the minimum distance between points in the cluster (to control how tightly packed the clusters are). For all UMAP spaces, we set the minimum distance to zero (for the tightest-packed clusters possible). We set the number of neighbors to be between 20 and 80. Changes with this hyperparameter within a reasonable range (20 to 80) did not change the clustering results.

Principal Component Analysis (PCA)[22], k-means clustering, and dbscan were implemented using sklearn. Although PCA does not require any hyperparameters, a scree plot was used to determine the number of principal components to keep given a specified threshold of explained variance. The value of k (the number of clusters) in k-means was determined to be between 3 and

6 such that it appeared to qualitatively distinguish the original nano-XANES image while reasonably explaining the reduced space. The clustering approach dbscan uses the epsilon hyperparameter, which we set to be one for all UMAP spaces. We qualitatively checked that this epsilon value appropriately labelled the UMAP clusters by visualizing the UMAP space color-coded by the dbscan labels.

3. Results and Discussion

A two-dimensional display of our sample, colored by maximum XANES intensity (and thus identifying regions with the highest photon counts) is shown in Fig. 1, where background spectra are filtered out such that only the sample region is examined. This sample – and thus dataset – is the same as the one found in Pattammattel, et al.[17] Each "pixel" (150 nm wide) represents a processed XANES spectrum.

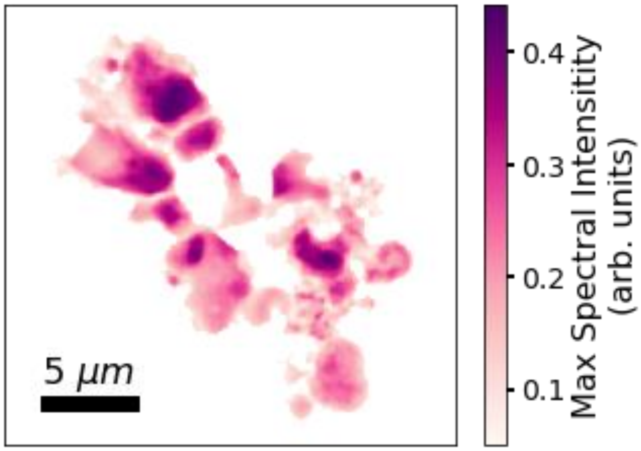


Figure 1 Nano-XANES map, color-coded by the maximum spectral intensity of the Fe K-edge XANES spectra (to indicate the most likely places with sample due to the high photon counts). Each pixel is 150 nm. Note that background spectra are filtered out.

Fig. 2 demonstrates our MPIS procedure in relation to both the standard nano-XANES analysis and LCF procedures. To start the MPIS procedure, we first apply Principal Component Analysis (PCA)[22] to the pre-processed ensemble of XANES spectra. Next, we have the option to encode multimodal data. Either we exclusively take the coefficients of the six highest principal components (determined by a scree plot, see Methods), or we use the joint encoding of those principal component coefficients with multimodal information.

To be specific, the multimodal encoding starts with an array of the principal component coefficients, i.e.,

$$\vec{S}(x,y) = (PC_1, \dots, PC_6) \tag{1}$$

Then, the spatial location and/or XRF of the four elements are appended to that array. In its most complex case, where both spatial location and XRF are jointly encoded, the encoding takes the form of the following vector:

$$\vec{S}(x,y) = (PC_1, \dots, PC_6, I_P^{XRF}, I_S^{XRF}, I_{Cr}^{XRF}, I_{Fe}^{XRF}, x, y) \tag{2}$$

where the first six components belong to the coefficients of the first six principal components, the next four components are the normalized XRF data (each of the P, S, and Cr XRF maps are divided by the Fe XRF map so that every pixel is normalized by total Fe fluorescence), and the last two components belong to the x and y positions (which are scaled to be between 0 and 1). The relative importance of the different components is then tuned by two new hyperparameters α and β dictating the informational strength or the importance of the XRF and spatial location, respectively. Thus, the above encoding is implemented as follows:

$$\vec{S}(x,y) = (PC_1, \dots, PC_6, \alpha I_P^{XRF}, \alpha I_S^{XRF}, \alpha I_{Cr}^{XRF}, \alpha I_{Fe}^{XRF}, \beta x, \beta y) \tag{3}$$

where α and β represent an independent scaling of importance for each distinct multimodal measurement. This procedure can be easily extended to encode other types of multimodal information.

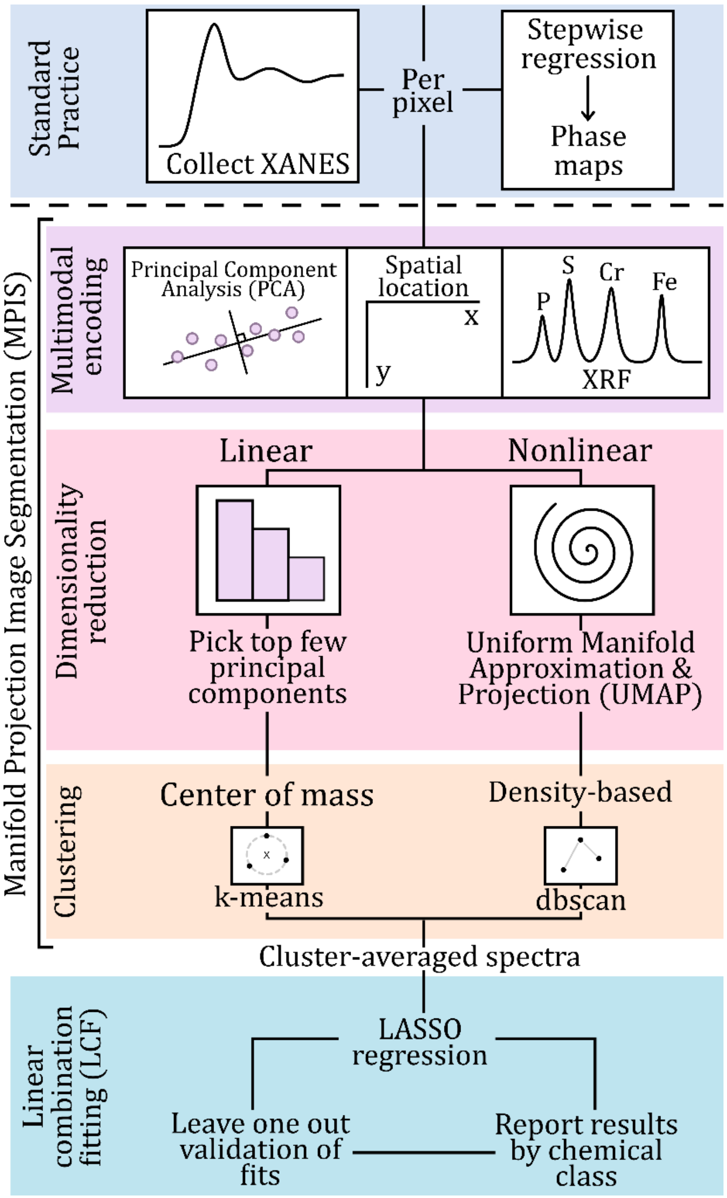

Standard Practice
Collect XANES
Per pixel
Stepwise regression
Phase maps
Manifold Projection Image Segmentation (MPIS)
Multimodal encoding
Principal Component Analysis (PCA)
Spatial location
x
y
S
Cr
Fe
P
XRF
Dimensionality reduction
Linear
Nonlinear
Pick top few principal components
Uniform Manifold Approximation & Projection (UMAP)
Clustering
Center of mass
Density-based
x
k-means
dbscan
Cluster-averaged spectra
Linear combination fitting (LCF)
LASSO regression
Leave one out validation of fits
Report results by chemical class

Figure 2 Our manifold projection image segmentation (MPIS) and linear combination fitting (LCF) pipeline for analyzing our nano-XANES image.

We then take the (multimodal) encoding and pass it to a nonlinear dimensionality reduction routine to identify spectral clusters. In general, applying dimensionality reduction before clustering increases the reliability of the clustering labels by combating the "curse of dimensionality" (as opposed to clustering applied directly to the spectra). We compared using a linear routine – namely PCA – to a nonlinear routine – namely Uniform Manifold Approximation and Projection (UMAP)[21] – when performing the dimensionality reduction step of MPIS. Prior work has shown that nonlinear dimensionality reduction, compared to linear, does better at disentangling the inherently nonlinear spectral features in X-ray absorption spectroscopy[23, 24], albeit linear routines are often sufficient[25, 26]. However, maintaining PCA as a preparation step for UMAP speeds up UMAP and filters out unimportant noise in the spectra, as shown in Fig. S1. Although we chose PCA as our linear routine, other linear methods such as non-negative matrix factorization (NMF)[27], could also be used.

While a center-of-mass-based clustering algorithm such as k-means[28] pairs well with PCA, we opted for a density-based cluster algorithm called dbscan[29] for the nonlinear embedding via UMAP. To see the effectiveness of UMAP and dbscan for clustering as opposed to PCA and k-means, see Fig. 3. The left panels in Fig. 3 show distinct and well separated clusters when UMAP and dbscan are used as opposed to overlapping or non-separated clusters using PCA and k-means Moreover, k-means needed five clusters to appropriately group the data in the PCA space, which is larger than the expected four known phases. Although Fig. 3 shows a two-dimensional projection of the data, we used six principal components in our MPIS pipeline as six principal components explained 97% of the variance of the data. See Figs. S2-5 for the triangle plots that visualize the PCA and UMAP hypercubes we used as well as other supplementary figures relating to MPIS.

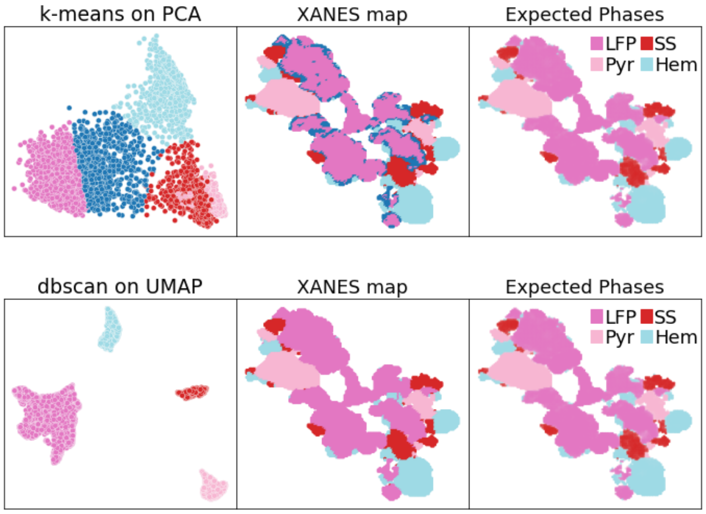


Figure 3 (top) k-means clustering on the first two principal components. (bottom) dbscan clustering on a two-dimensional UMAP embedding. The clusters and labeling in the two-dimensional UMAP representation not only match expectations, but they are easier to see and thus interpret than the k-means clusters on the top two PCA components.

Finally, we identified the composition of each cluster by performing linear combination fitting (LCF) to a reference library using the MPIS cluster-averaged spectra. To do so, we utilized the procedure presented in Jahrman, et al.[14], involving Least Absolute Selection and Shrinkage Operator (LASSO) regression instead of stepwise regression. However, instead of bootstrapping our data to generate estimates in uncertainty, we utilized leave-one-out validation. Specifically, we refit each spectrum with one reference spectrum in the library removed at a time and noted the changes in the fit results. In addition to speeding up computation time by avoiding stepwise regression, we reduced the total number of fits by performing LCF on the average spectrum for each cluster rather than the spectrum at every pixel individually. See the Supplementary Information for details on LASSO regression. In brief, hyperparameters were chosen via 5-fold cross validation.

Our reference library for LCF was composed of both the known phases – LiFe(II)PO_4 (LFP), pyrite, stainless steel, and hematite – and additional mineral phases to model a typical experiment, namely HFO (hydrous ferric oxyhydroxide), goethite, maghemite, magnetite, Fe_3P, Fe(III)PO_4, and Fe(III)SO_4[12]. Specifically, hematite, goethite, maghemite, and HFO are all oxides and have very similar spectra, while Fe_3P has the same oxidation state as elemental Fe, which is the same as stainless steel. The selection of this library was based on a quick XRF measurement and the availability of experimental reference spectra. Moreover, this reference library represents a realistic uncertainty for chemical speciation of Fe-phases in heterogeneous samples with *a priori* knowledge.

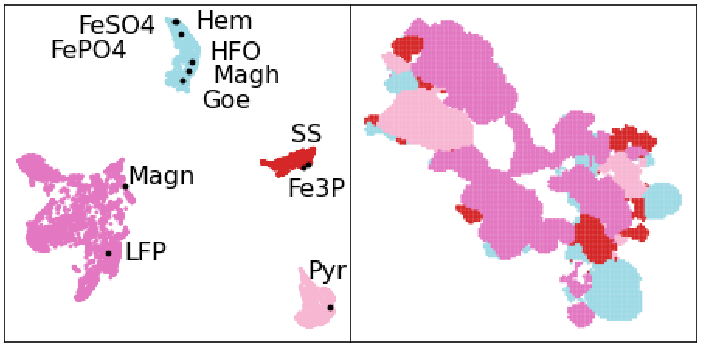


Figure 4 Reference chemical classes. Often, LCF results are reported using the chemical class of the references. These classes are usually created using chemical knowledge of the system. Instead, we offer a completely data-driven way one can generate these classes, specifically by projecting references onto the UMAP space determined by the experimental spectra.

Finally, we reported LCF fit results by the chemical classes for the references, which we developed by projecting the reference spectra onto the UMAP space of the experimental data, as shown in Fig. 4. We divided the references from the same cluster if the XRF would be able to distinguish between references. For example, while Fe_3P and stainless steel appeared in the same cluster, they are theoretically distinguishable using both the P and Cr XRF data. Following this procedure, all references were split into their own class besides the “oxides” – hematite, maghemite, HFO, and goethite – which were grouped into one combined class. To see the correlation matrix for references, see Fig. S6.

We hypothesized that applying MPIS and LASSO regression rather than pixel-by-pixel stepwise regression would speed up computation time while maintaining accuracy. We ran both procedures and found MPIS took about 30 seconds compared to the standard pixel-by-pixel stepwise regression procedure of enumerating all quaternary combinations of 11 references, which took 4 minutes (using 8 GB RAM on a 2-core Intel i5 CPU). However, the time complexity of stepwise regression fits grows as $O(n^k)$ given the reference library size n and the combination size k, so a larger reference library will greatly increase computation time.

We then compared the effect of encoding the XRF and spatial location of every spectrum into MPIS on the LCF results, as shown in Fig 5. The uppermost left panel shows the LCF results with no multimodal encoding in MPIS. We compared predicted coefficients using the “standard” approach (non-negative least squares for every pixel), likewise using “LASSO” for every pixel, and then using “MPIS” and predicting concentrations from cluster averages. These coefficients were scored against the “true” concentrations, which were obtained by non-negative least squares per pixel using only the four known phases in the reference library (rather than all 11 as in the “standard” case).

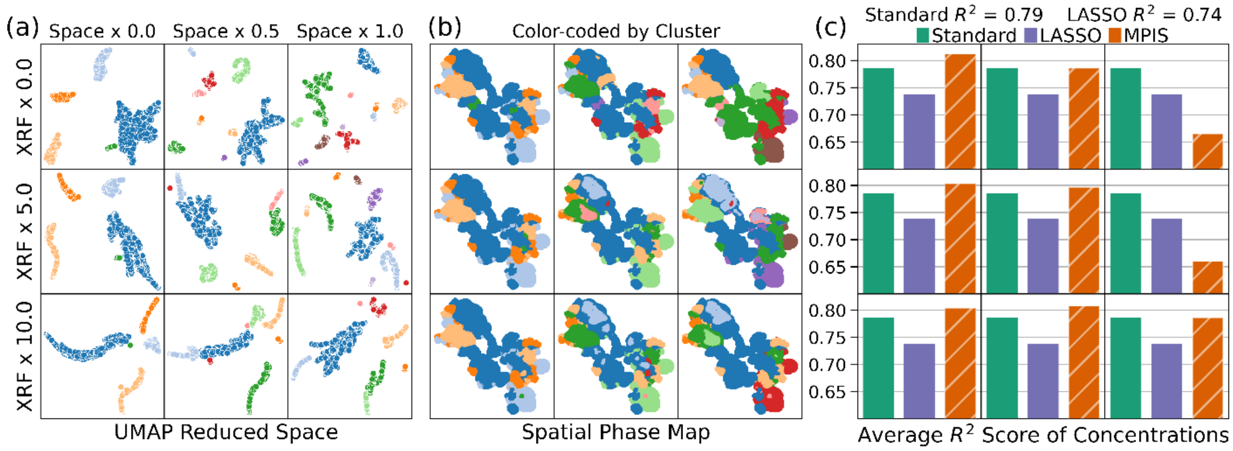


Figure 5 (a) Effects on the clusters when encoding XRF data and spatial location into the MPIS pipeline. (b) The resulting 2D phase maps, colored by cluster. (c) Score of linear combination fitting (LCF) predictions via the standard pixel-by-pixel analysis ("Standard"), pixel-by-pixel LASSO regression ("LASSO"), and LASSO regression via MPIS ("MPIS"). The upper leftmost panel shows no joint information encoding.

In general, MPIS scored just as well as the standard approach (if the spatial strength is not too large) but in less time. Moreover, by identifying domains, fits are less sensitive to uncorrelated noise in the dataset. To demonstrate this effect, we augmented the experimental spectra with additional uncorrelated Gaussian noise with increasing intensity and compared domain identification using the standard pixel-by-pixel analysis with domains identified by MPIS, as shown in Fig. 6. Because the standard procedure constructs phase maps after LCF, there are spurious single-pixel phases when the noise is large. However, MPIS is more robust against these fluctuations. Moreover, generating cluster-averaged spectra via MPIS is an informed way to average noisy spectra together for LCF fits without needing to lose resolution by Gaussian blurring the image. Furthermore, when noise is so high that MPIS on just the spectra fails, encoding XRF and spatial location into MPIS recovers the analysis, as shown in Fig. S7. Fig. S8 compares the average error in predicting concentrations.

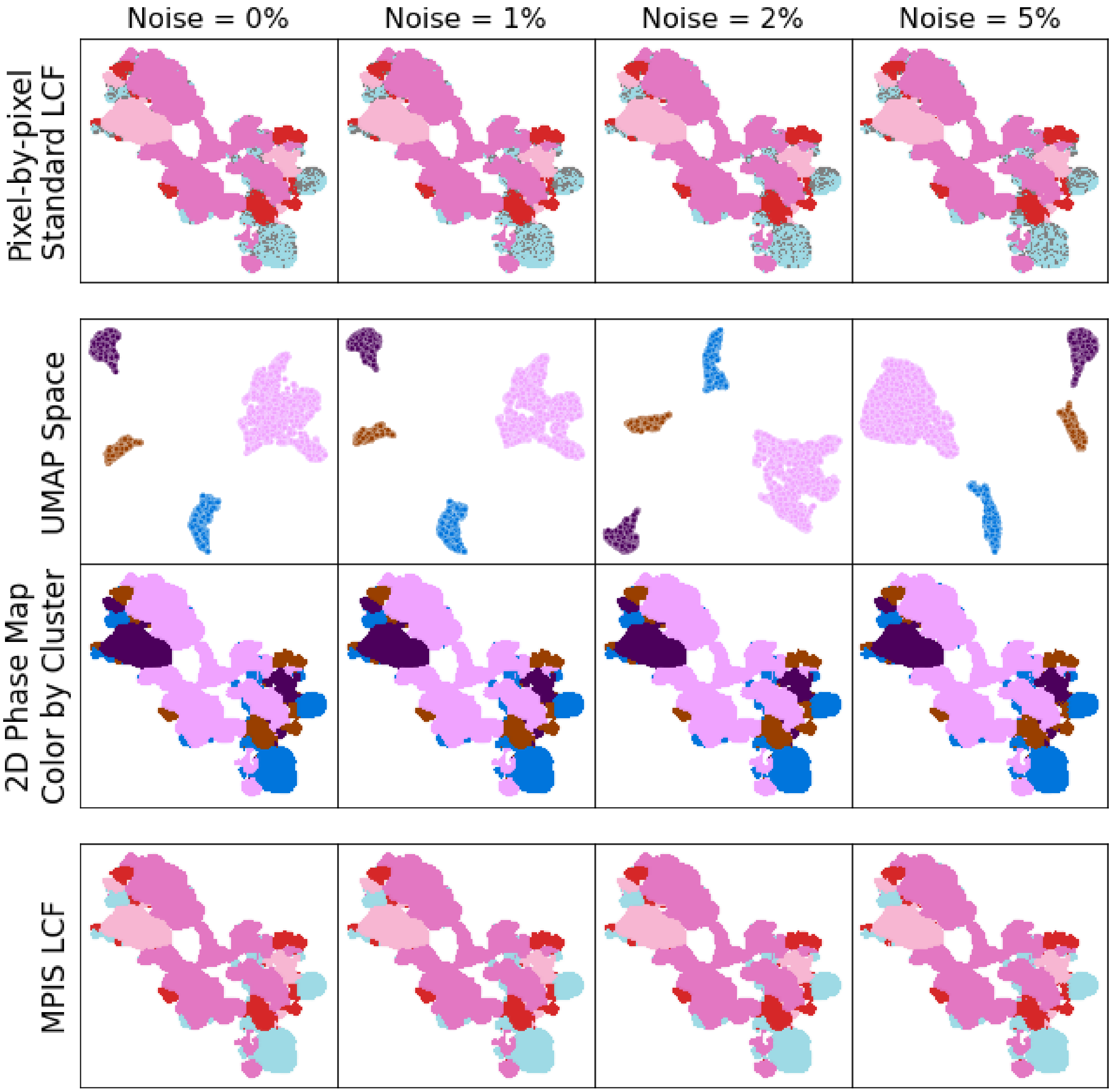


Figure 6 Adding noise (as a percentage of spectral intensity) to the experimental spectra causes pixel-by-pixel analysis to have small unphysical fluctuations in the phase maps, resulting from

uncertain LCF fits (top row). By applying MPIS (second row), the phase maps (third row) are more robust to noise, demonstrated by the consistent LCF results (bottom row).

5. Conclusions

The standard procedure for analyzing nano-XANES imaging is performing linear combination fitting (LCF) via stepwise regression for every pixel independently and then constructing phase maps using the fit results. Instead of stepwise regression, we encourage sparsity by performing LASSO regression to LCF onto our reference library. LASSO regression reduces computation time by decreasing the required number of fits. We also implement manifold projection image segmentation (MPIS) to cluster experimental spectra first before performing LCF, enabling LCF to only identify the composition of the phases rather than informing the spatial maps. By identifying domains first, we decouple the reliance on correct LCF fit results for appropriate phase maps, which can be greatly impacted by noise. The other benefit to using MPIS rather than the traditional deep learning approaches is that it requires fewer hyperparameters (which avoids overfitting), all of which are physically motivated, and can be more computationally efficient. Moreover, MPIS is adaptable to include multimodal data, which we demonstrated by encoding X-ray Fluorescence and spatial location of pixels in addition to the XANES spectra. Because the spatial location of pixels is encoded as additional information, the basic procedure of MPIS can be applied to any ensemble-based spectroscopy measurements where clustering is important. Furthermore, we propose that MPIS can be applied to any spectral imaging measurement, such as Scanning Transmission X-ray Microscopy (STXM), where the experimental resolution (pixel size) is smaller than the intrinsic length scale of domains.

6. Acknowledgements

We thank M. Marcus and B. Toner for their beneficial discussions regarding STXM and sharing their STXM datasets with us. This work is supported by the U.S. Department of Energy (DOE) Office of Science Scientific User Facilities project titled "Integrated Platform for Multimodal Data Capture, Exploration and Discovery Driven by AI Tools". M.K.Y.C. acknowledges the support from the BES SUFD Early Career award. This research used Hard X-ray Nanoprobe beamline of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. Work performed at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. All code is available through GitHub at github.com/stetef/nano-XANES-microscopy-of-Fe[30].

7. References

(1) Zhang, C.; Zhou, J.; Wang, H.; Tan, T.; Cui, M.; Huang, Z.; Wang, P.; Zhang, L. Multi-Species Individual Tree Segmentation and Identification Based on Improved Mask R-CNN and UAV Imagery in Mixed Forests. *Remote Sensing* **2022**, *14* (4), 874. DOI: 10.3390/rs14040874.

(2) Jakubowski, M. K.; Li, W.; Guo, Q.; Kelly, M. Delineating Individual Trees from Lidar Data: A Comparison of Vector- and Raster-based Segmentation Approaches. *Remote Sensing* **2013**, *5* (9), 4163-4186. DOI: 10.3390/rs5094163.

(3) Yapp, C.; Novikov, E.; Jang, W.-D.; Vallius, T.; Chen, Y.-A.; Cicconet, M.; Maliga, Z.; Jacobson, C. A.; Wei, D.; Santagata, S.; et al. UnMICST: Deep learning with real augmentation for robust segmentation of highly multiplexed images of human tissues. *Communications Biology* **2022**, *5* (1), 1263. DOI: 10.1038/s42003-022-04076-3.

(4) Schwartzkopf, W. C.; Bovik, A. C.; Evans, B. L. Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images. *IEEE Transactions on Medical Imaging* **2005**, *24* (12), 1593-1610. DOI: 10.1109/TMI.2005.859207.

(5) Hu, H.; Yin, R.; Brown, H. M.; Laskin, J. Spatial Segmentation of Mass Spectrometry Imaging Data by Combining Multivariate Clustering and Univariate Thresholding. *Analytical Chemistry* **2021**, *93* (7), 3477-3485. DOI: 10.1021/acs.analchem.0c04798.

(6) Stolarek, I.; Samelak-Czajka, A.; Figlerowicz, M.; Jackowiak, P. Dimensionality reduction by UMAP for visualizing and aiding in classification of imaging flow cytometry data. *iScience* **2022**, *25* (10), 105142. DOI: 10.1016/j.isci.2022.105142 (acccessed 2023/02/28).

(7) Nakai, I.; Numako, C.; Hayakawa, S.; Tsuchiyama, A. Chemical speciation of geological samples by micro-XANES techniques. *Journal of Trace and Microprobe Techniques* **1998**, *16* (1), 87-98.

(8) Belissont, R.; Munoz, M.; Boiron, M. C.; Luais, B.; Mathon, O. Germanium Crystal Chemistry in Cu-Bearing Sulfides from Micro-XRF Mapping and Micro-XANES Spectroscopy. *Minerals* **2019**, *9* (4), 227. DOI: 10.3390/min9040227.

(9) Cusack, M.; Dauphin, Y.; Cuif, J. P.; Salome, M.; Freer, A.; Yin, H. Micro-XANES mapping of sulphur and its association with magnesium and phosphorus in the shell of the brachiopod, Terebratulina retusa. *Chemical Geology* **2008**, *253* (3-4), 172-179. DOI: 10.1016/j.chemgeo.2008.05.007.

(10) Bonnin-Mosbah, M.; Métrich, N.; Susini, J.; Salomé, M.; Massare, D.; Menez, B. Micro X-ray absorption near edge structure at the sulfur and iron K-edges in natural silicate glasses. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2002**, *57* (4), 711-725. DOI: https://doi.org/10.1016/S0584-8547(01)00407-4.

(11) Mino, L.; Borfecchia, E.; Groppo, C.; Castelli, D.; Martinez-Criado, G.; Spiess, R.; Lamberti, C. Iron oxidation state variations in zoned micro-crystals measured using micro-XANES. *Catalysis Today* **2014**, *229*, 72-79. DOI: 10.1016/j.cattod.2013.11.002.

(12) Pattammattel, A.; Tappero, R.; Ge, M.; Chu, Y. S.; Huang, X.; Gao, Y.; Yan, H. High-sensitivity nanoscale chemical imaging with hard x-ray nano-XANES. *Science Advances* **2020**, *6* (37), eabb3615. DOI: 10.1126/sciadv.abb3615 (acccessed 2022/10/20).

(13) Bunker, G. *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*; Cambridge University Press, 2010. DOI: DOI: 10.1017/CBO9780511809194.

(14) Jahrman, E. P.; Yu, L. L.; Krekelberg, W. P.; Sheen, D. A.; Allison, T. C.; Molloy, J. L. Assessing arsenic species in foods using regularized linear regression of the arsenic K-edge X-ray absorption near edge structure. *Journal of Analytical Atomic Spectrometry* **2022**, *37* (6), 1247-1258, 10.1039/D1JA00445J. DOI: 10.1039/D1JA00445J.

(15) Nazaretski, E.; Yan, H.; Lauer, K.; Bouet, N.; Huang, X.; Xu, W.; Zhou, J.; Shu, D.; Hwu, Y.; Chu, Y. S. Design and performance of an X-ray scanning microscope at the Hard X-ray Nanoprobe beamline of NSLS-II. *Journal of Synchrotron Radiation* **2017**, *24* (6), 1113-1119.
(16) Yan, H.; Bouet, N.; Zhou, J.; Huang, X.; Nazaretski, E.; Xu, W.; Cocco, A. P.; Chiu, W. K. S.; Brinkman, K. S.; Chu, Y. S. Multimodal hard x-ray imaging with resolution approaching 10 nm for studies in material science. *Nano Futures* **2018**, *2* (1), 011001. DOI: 10.1088/2399-1984/aab25d.
(17) Pattammattel, A.; Tappero, R.; Gavrilov, D.; Zhang, H.; Aronstein, P.; Forman, H. J.; O'Day, P. A.; Yan, H.; Chu, Y. S. Multimodal X-ray nano-spectromicroscopy analysis of chemically heterogeneous systems. *Metallomics* **2022**, *14* (10), mfac078. DOI: 10.1093/mtomcs/mfac078 (acccessed 10/20/2022).
(18) Thevenaz, P.; Ruttimann, U. E.; Unser, M. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing* **1998**, *7* (1), 27-41. DOI: 10.1109/83.650848.
(19) Ravel, B.; Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *Journal of Synchrotron Radiation* **2005**, *12* (4), 537-541, https://doi.org/10.1107/S0909049505012719. DOI: https://doi.org/10.1107/S0909049505012719 (acccessed 2021/04/22).
(20) Li, L.; Hanfei, Y.; Wei, X.; Dantong, Y.; Annie, H.; Wah-Keat, L.; Stuart, I. C.; Yong, S. C. PyXRF: Python-based X-ray fluorescence analysis package. In *Proc.SPIE*, 2017; X-Ray Nanoimaging: Instruments and Methods III: Vol. 10389, p 103890U. DOI: 10.1117/12.2272585.
(21) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426* **2020**, (1802.03426).
(22) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2* (1), 37-52. DOI: https://doi.org/10.1016/0169-7439(87)80084-9.
(23) Tetef, S.; Govind, N.; Seidler, G. T. Unsupervised machine learning for unbiased chemical classification in X-ray absorption spectroscopy and X-ray emission spectroscopy. *Phys. Chem. Chem. Phys.* **2021**, *23* (41), 23586-23601, 10.1039/D1CP02903G. DOI: 10.1039/D1CP02903G.
(24) Tetef, S.; Kashyap, V.; Holden, W. M.; Velian, A.; Govind, N.; Seidler, G. T. Informed Chemical Classification of Organophosphorus Compounds via Unsupervised Machine Learning of X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy. *The Journal of Physical Chemistry A* **2022**, *126* (29), 4862-4872. DOI: 10.1021/acs.jpca.2c03635.
(25) Lerotic, M.; Jacobsen, C.; Gillow, J. B.; Francis, A. J.; Wirick, S.; Vogt, S.; Maser, J. Cluster analysis in soft X-ray spectromicroscopy: Finding the patterns in complex specimens. *Journal of Electron Spectroscopy and Related Phenomena* **2005**, *144-147*, 1137-1143. DOI: https://doi.org/10.1016/j.elspec.2005.01.158.
(26) Marcus, M. A. Data analysis in spectroscopic STXM. *Journal of Electron Spectroscopy and Related Phenomena* **2023**, *264*, 147310. DOI: https://doi.org/10.1016/j.elspec.2023.147310.
(27) Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401* (6755), 788-791. DOI: 10.1038/44565.
(28) Sinaga, K. P.; Yang, M. S. Unsupervised K-Means Clustering Algorithm. *IEEE Access* **2020**, *8*, 80716-80727. DOI: 10.1109/ACCESS.2020.2988796.
(29) Hahsler, M.; Piekenbrock, M.; Doran, D. dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software* **2019**, *91* (1), 1 - 30. DOI: 10.18637/jss.v091.i01 (acccessed 2021/12/12).

(30) *github.com/stetef/nano-XANES-microscopy-of-Fe*. (accessed 2023 July).

Supplementary Information

Manifold Projection Image Segmentation for Nano-XANES Imaging

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

Linear combination fitting objective function

Our linear combination fitting (LCF) objective function is:

$$\hat{\vec{c_j}} = argmin_{\vec{c_j}} \left[\frac{1}{2} \left|\left|\vec{y_j} - R^T \cdot \vec{c_j}\right|\right|_2^2 + \lambda_1 \left|\left|\vec{c_j}\right|\right|_1 + \lambda_2 \left|\left|1 - \Sigma_i c_{ij}\right|\right|_2^2 \right]$$

where **y** is the unknown experimental spectra, R is the matrix composed of reference spectra, and **c** is the coefficients contributing to the spectra. The first term represents reconstruction error (via a L_2 norm, which is equivalent to a Euclidean distance metric), and the second term is the regularizer, modified by a Lagrange multiplier. The regularization was set to be the L1 norm to encourage sparsity. These terms effectively constitute LASSO regression. However, the input for each spectrum is the same – specifically the reference set R – so each spectrum is fit independently of the others.

The hyperparameters for the fits were found via 5-fold cross-validation on a dataset composed of linear combinations of reference spectra (with forced sparsity and various levels of noise introduced). Specifically, we found a λ_1 value of 0.0006 and λ_2 value of 10 to be best, with consistent convergence onto a solution. Again, note that this objective function is minimized for every spectrum (or data point) and is therefore not trained on any training dataset as the input (the reference set R) is the same for every fit. To minimize the above objective function, we used `scipy`'s minimize function with bounds on the weights to be in the range [30]. Moreover, we used `scipy`'s built-in Sequential Least Squares Programming (SLSQP) optimization method, which is a quasi-Newton method.

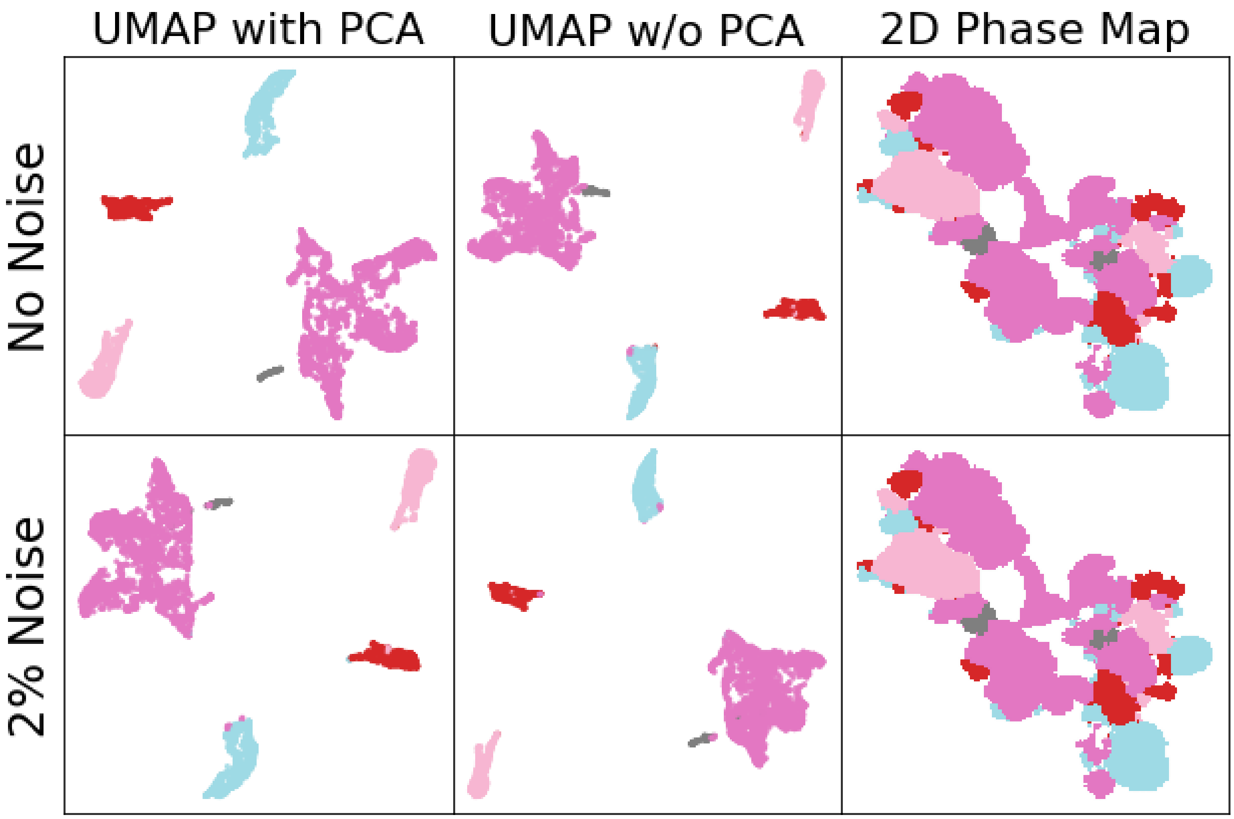


Fig. S1 UMAP spaces with and without PCA processing. UMAP applied to the first 6 principal components produces clusters that are very similar to the clusters made when UMAP is applied directly to the spectra, both on the raw experimental data and with augmented noise added to the spectra.

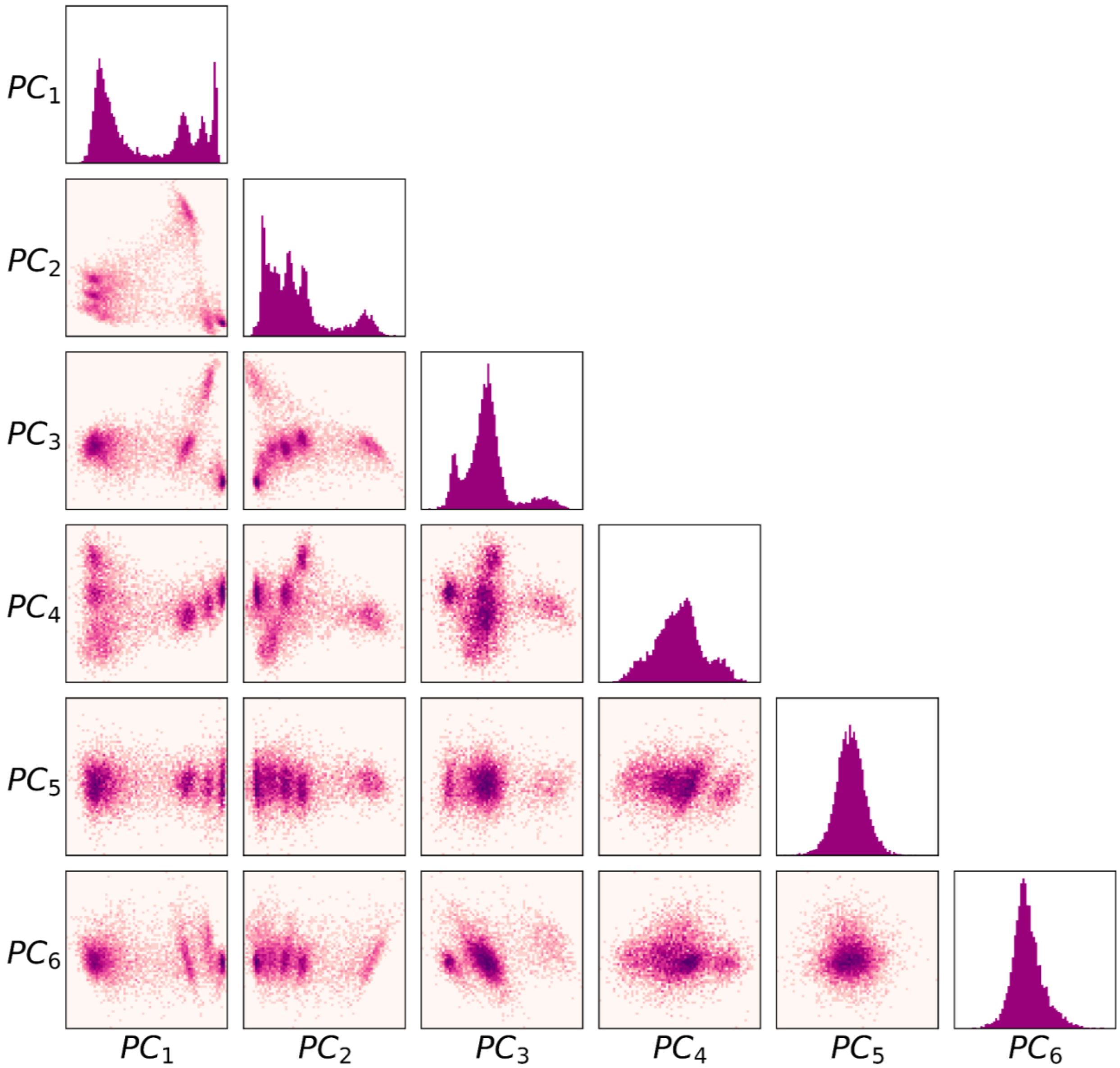


Fig. S2 PCA triangle plot of all two-dimensional projects of the six-dimensional hypercube of the top principal components of the spectral dataset. Six dimensions were chosen because it takes the top six principal components (PCs) to explain 97% of the variance.

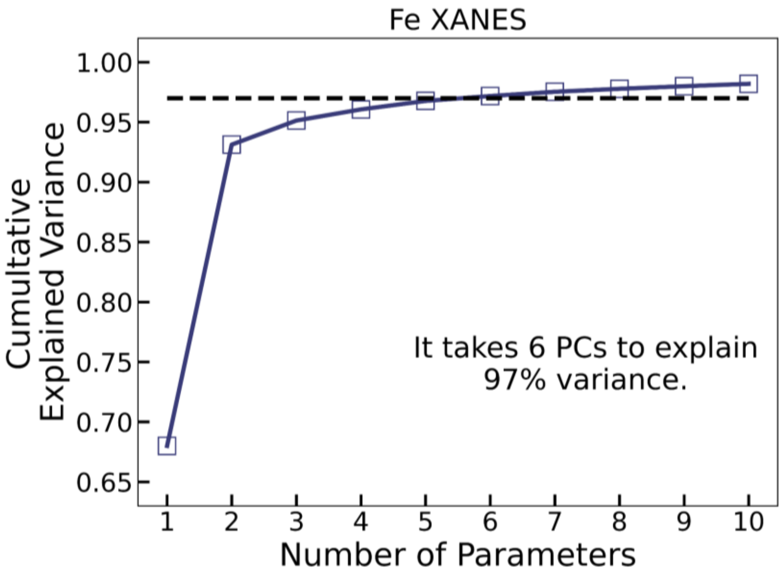


Fig. S3 Scree plot of experimental spectra. It takes 6 PCs to explain 97% variance (dashed line).

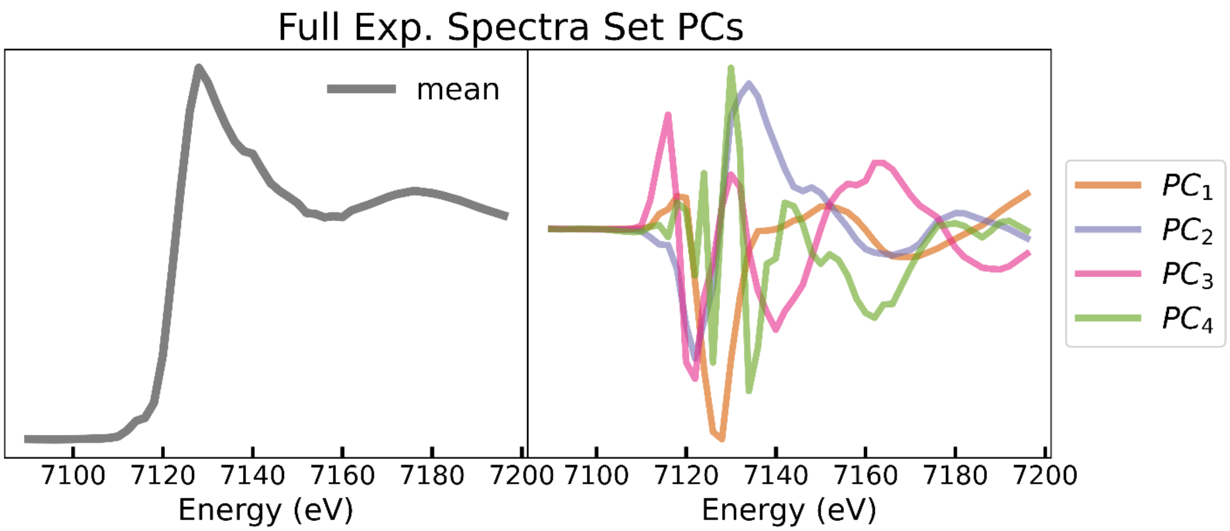


Fig. S4 First four principal components of the measured spectra.

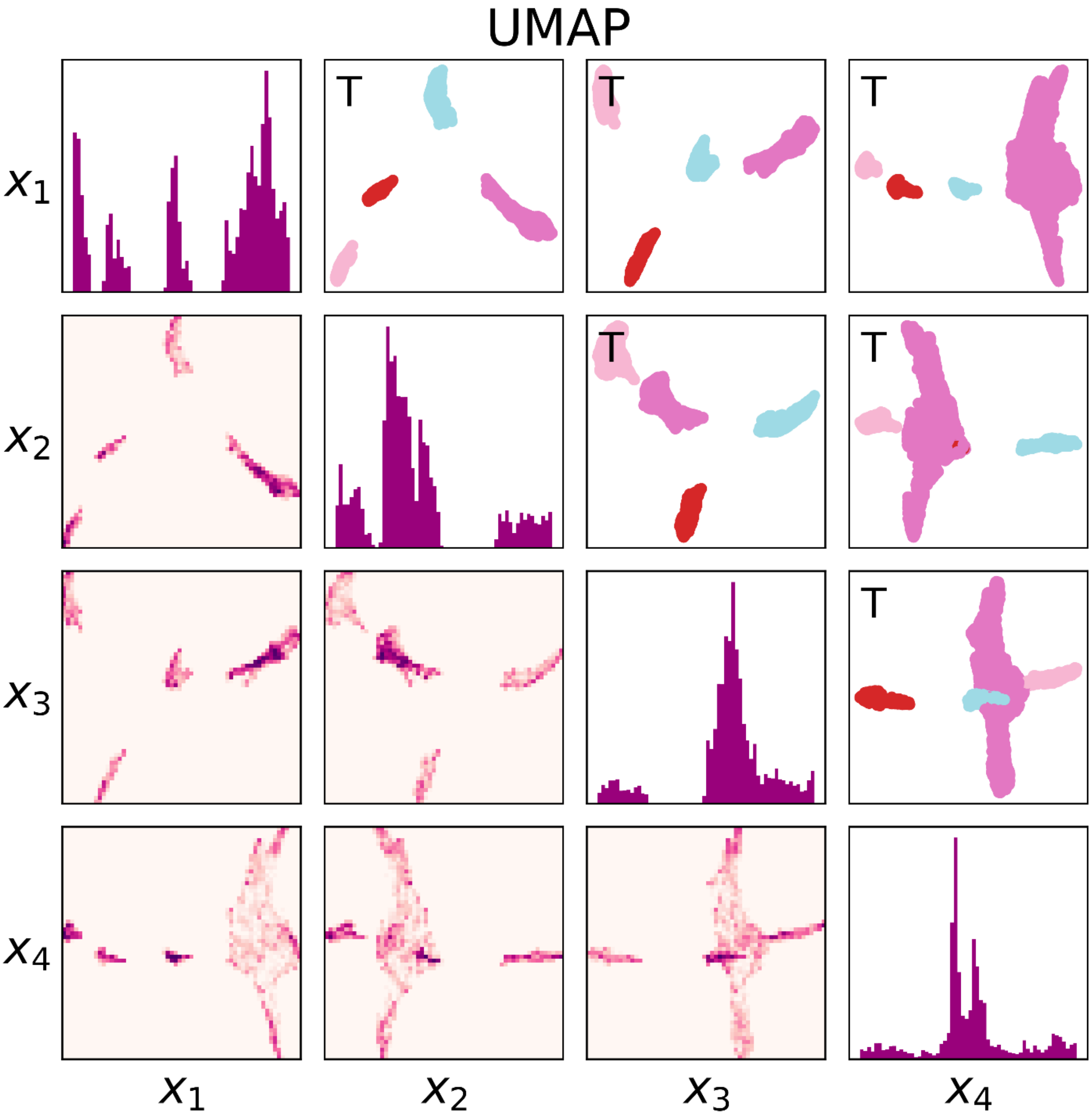


Fig. S5 Four-dimensional UMAP hypercube (applied to the top PCA components), with two-dimensional projections (color-coded by density) shown in the bottom left corner. The upper right corner is composed of the same projections (transposed so that the upper and lower triangles match), except instead color-coded by the dbscan clustering labels.

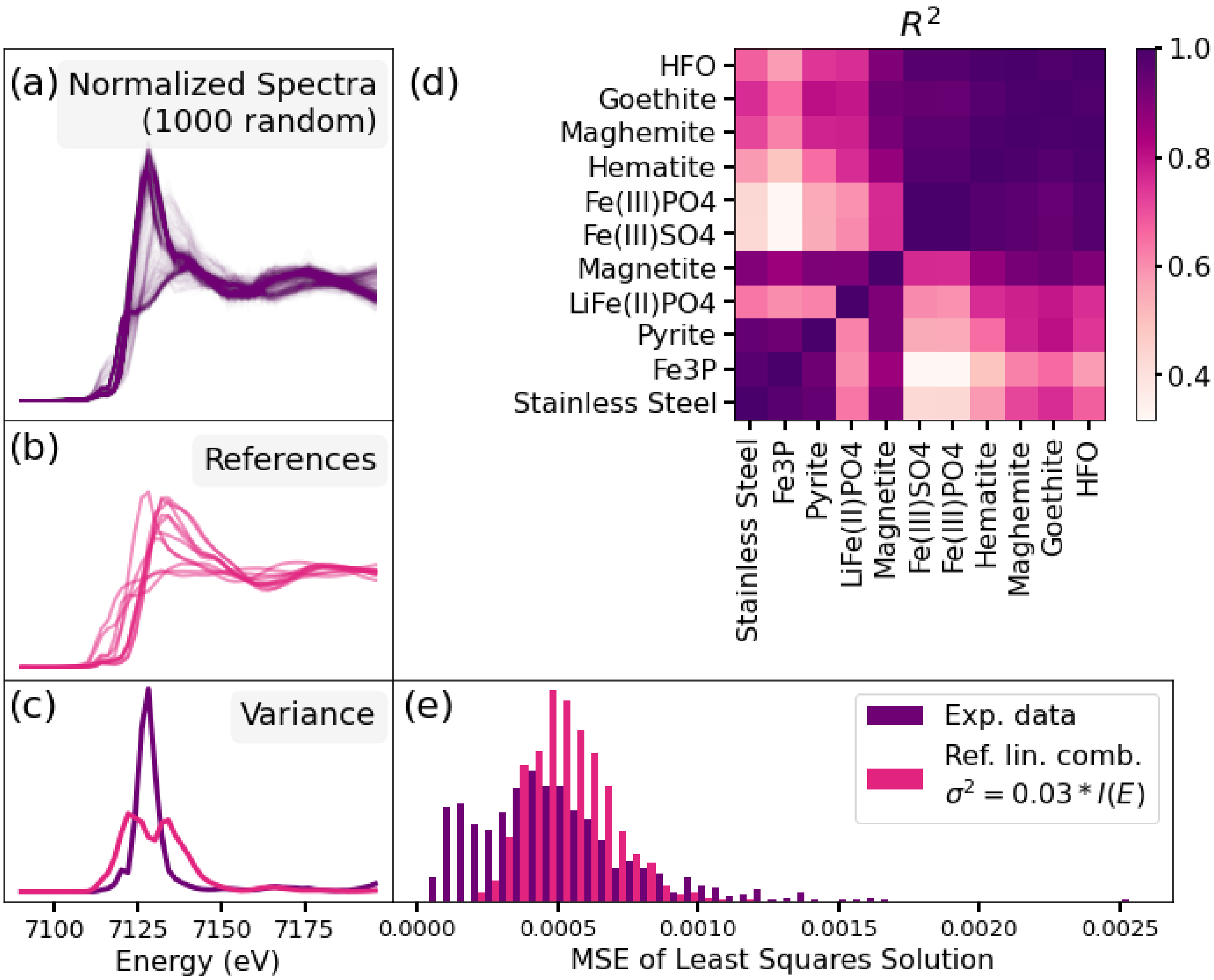


Fig. S6 Correlation of reference spectra. (a) 1000 randomly sampled experimental spectra (which passed the background filter) normalized following the standard procedure in Athena [19]. (b) Reference spectra used in this study (11 total). (c) Variation of both the 1000 experimental spectra and the reference spectra. (e) Mean squared error between the original spectra and the fitted (via least squares) spectra of both the experimental data and true linear combinations of references (with random normal noise with a variance of 3% of the spectral intensity to model true experimental noise).

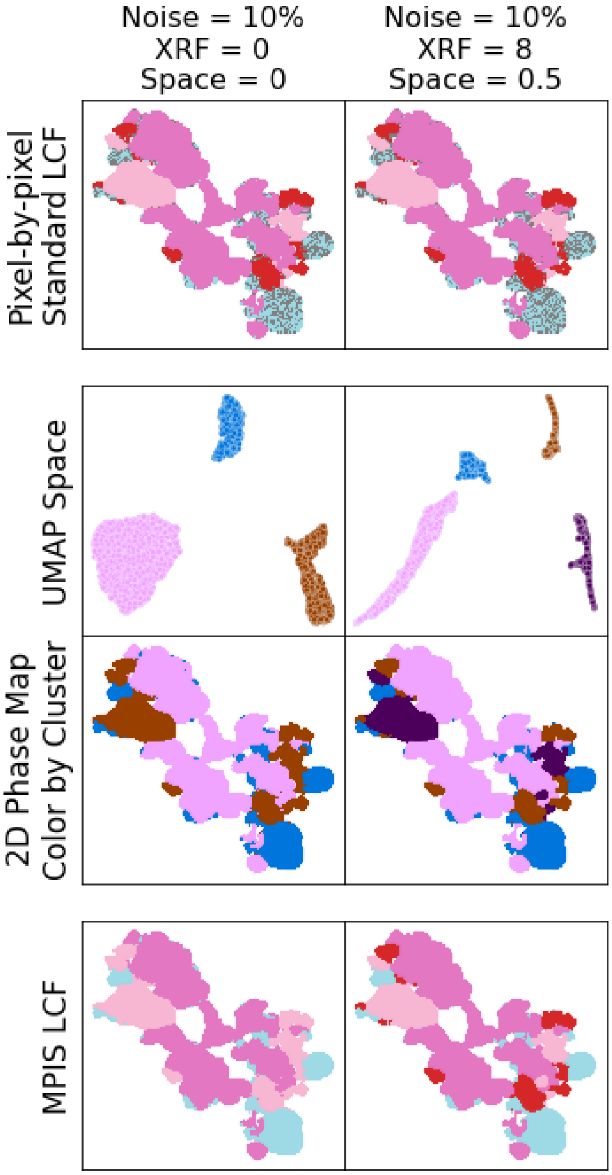


Fig. S7 Augmented MPIS versus noise. When noise is set to 10%, only three clusters appear (left column). However, adding XRF and spatial encoding recovers the lost cluster (right column).

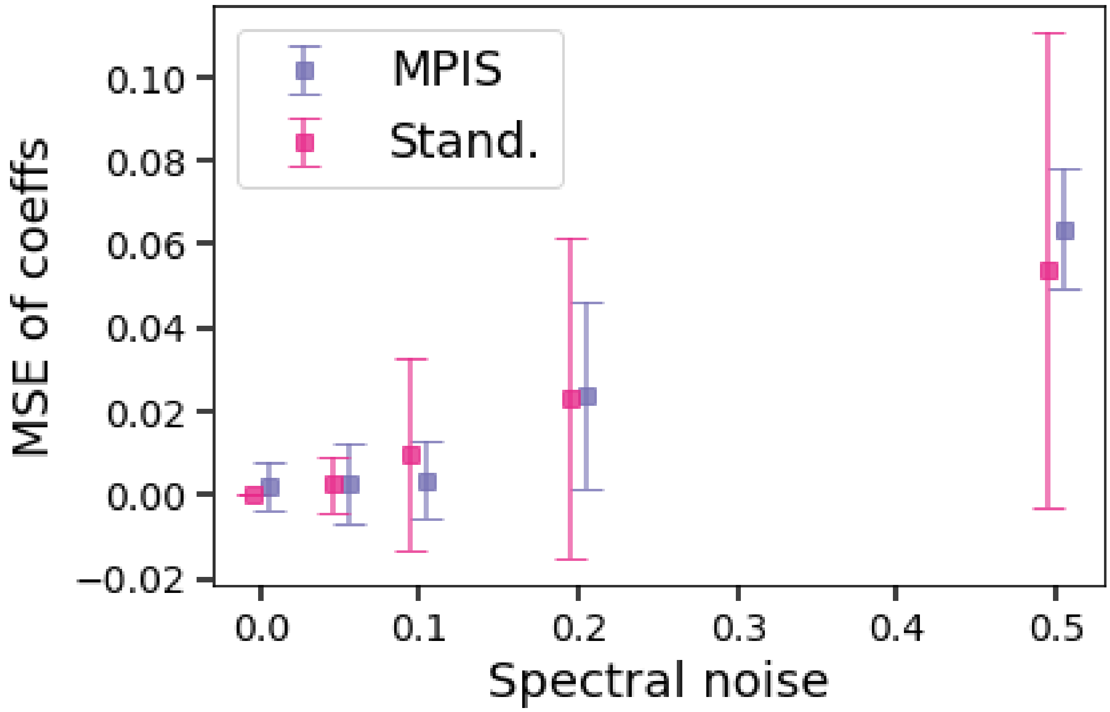


Fig. S8 Error in predicted concentrations versus noise (on a dataset composed of true linear combinations of references) for both our MPIS cluster-averaged pipeline and the standard individual spectrum analysis. The variance in predictions decreases when spectra are averaged together in an informed way via MPIS.