SIGMA: Spectral Interpretation using Gaussian Mixtures and Autoencoder

Po-Yen Tung[1], Hassan Aftab Sheikh[1], Matthew R Ball[1], Farhang Nabiei[2], and Richard Harrison[1]

[1]University of Cambridge
[2]MediaTek Research

November 24, 2022

Abstract

Identification of unknown micro- and nano-sized mineral phases is commonly achieved by analyzing chemical maps generated from hyperspectral imaging datasets, particularly scanning electron microscope - energy dispersive X-ray spectroscopy (SEM-EDS). However, the accuracy and reliability of mineral identification are often limited by subjective human interpretation, non-ideal sample preparation, and the presence of mixed chemical signals generated within the electron-beam interaction volume. Machine learning has emerged as a powerful tool to overcome these problems. Here, we propose a machine-learning approach to identify unknown phases and unmix their overlapped chemical signals. This approach leverages the guidance of Gaussian mixture modeling clustering fitted on an informative latent space of pixel-wise elemental datapoints modeled using a neural network autoencoder, and unmixes the overlapped chemical signals of phases using non-negative matrix factorization. We evaluate the reliability and the accuracy of the new approach using two SEM-EDS datasets: a synthetic mixture sample and a real-world particulate matter sample. In the former, the proposed approach successfully identifies all major phases and extracts background-subtracted single-phase chemical signals. The unmixed chemical spectra show an average similarity of 83.0% with the ground truth spectra. In the second case, the approach demonstrates the ability to identify potentially magnetic Fe-bearing particles and their background-subtracted chemical signals. We demonstrate a robust approach that brings a significant improvement to mineralogical and chemical analysis in a fully automated manner. The proposed analysis process has been built into a user-friendly Python code with a graphical user interface for ease of use by general users.

Hosted file

`essoar.10511396.1.docx` available at https://authorea.com/users/551607/articles/604431-sigma-spectral-interpretation-using-gaussian-mixtures-and-autoencoder

Po-Yen Tung[1,2*], Hassan A. Sheikh[1], Matthew Ball[1], Farhang Nabiei[1,3], Richard J. Harrison[1]

[1]Department of Earth Sciences, University of Cambridge, Cambridge, UK

[2]Department of Materials Science and Metallurgy, University of Cambridge, Cambridge, UK

[3]MediaTek Research, Cambourne, UK

*Corresponding author: Po-Yen Tung (pyt21@cam.ac.uk)

Key Points:

- We develop a machine-learning approach to automatically identify unknown mineral phases and unmix their compositional spectra.
- The approach successfully identifies all phases in a synthetic mixture dataset with an average spectrum similarity 83% to the ground truth.
- The approach demonstrates its robustness by identifying unknown Fe-bearing particles and background-subtracted chemical signals.

Abstract

Identification of unknown micro- and nano-sized mineral phases is commonly achieved by analyzing chemical maps generated from hyperspectral imaging datasets, particularly scanning electron microscope - energy dispersive X-ray spectroscopy (SEM-EDS). However, the accuracy and reliability of mineral identification are often limited by subjective human interpretation, non-ideal sample preparation, and the presence of mixed chemical signals generated within the electron-beam interaction volume. Machine learning has emerged as a powerful tool to overcome these problems. Here, we propose a machine-learning approach to identify unknown phases and unmix their overlapped chemical signals. This approach leverages the guidance of Gaussian mixture modeling clustering fitted on an informative latent space of pixel-wise elemental datapoints modeled using a neural network autoencoder, and unmixes the overlapped chemical signals of phases using non-negative matrix factorization. We evaluate the reliability and the accuracy of the new approach using two SEM-EDS datasets: a synthetic mixture sample and a real-world particulate matter sample. In the former, the proposed approach successfully identifies all major phases and extracts background-subtracted single-phase chemical signals. The unmixed chemical spectra show an average similarity of 83.0% with the ground truth spectra. In the second case, the approach demonstrates the ability to identify potentially magnetic Fe-bearing particles and their background-subtracted chemical signals. We demonstrate a robust approach that brings a significant improvement to mineralogical and chemical analysis in a fully automated manner. The proposed analysis process has been built into a user-friendly Python code with a graphical user interface for ease of use by general users.

1 Introduction

Hyperspectral imaging (HSI) data is a two-dimensional pixelated dataset, where each pixel stores a one-dimensional array of spectral data, forming a three-dimensional datacube. HSI data provides vast quantities of spatial and spectral information and has been widely applied in various fields, such as remote sensing (*Blackburn*, 2006), vegetation and water source control (*Adam et al.*, 2010; *Govender et al.*, 2007), food safety (*Carrasco et al.*, 2003; *Feng & Sun*, 2012; *Gowen et al.*, 2007), and biomedical sciences (*Afromowitz et al.*, 1988; *Carrasco et al.*, 2003; *Gendrin et al.*, 2008). In mineral sciences, scanning electron microscopy (SEM) is one of the most used microanalysis techniques. SEM provides measurements of surface morphology (by the detection of secondary electrons), elemental composition (by X-ray spectroscopy), crystallography (by backscattered electrons), chemical bonding (by Auger electrons), and electronic state (by cathodoluminescence) (*Goldstein et al.*, 2017; *Zaefferer & Habler*, 2017). X-ray emission can be analyzed by energy dispersive X-ray spectroscopy (EDS), where an X-ray spectrum is recorded for each pixel scanned by an electron beam over the sample surface, building up an HSI dataset. HSI-EDS data is frequently used for chemical phase identification. By integrating over manually defined intervals of the EDS spectra for each pixel, elemental distribution maps are generated in qualitative and quantitative manners. Traditionally, phase identification is conducted by analyzing the elemental maps superimposed on morphological SEM images "by hand". However, this process is time-consuming and prone to error, particularly for large datasets. Furthermore, the resulting qualitative information only relies on subjective human interpretation, reducing the reliability and reproducibility, particularly when dealing with unknown samples. Automating this process with high accuracy and reliability is critical for studying natural materials.

Multivariate statistical analysis (MSA) is a popular choice for automated solutions (*Bosman et al.*, 2006; *Kannan et al.*, 2018; *Kotula et al.*, 2003; *Malinowski & Howery*, 1980; *Teng & Gauvin*, 2020). Principal component analysis (PCA) and non-negative matrix factorization (NMF) are two widely used MSA algorithms for the exploration of the HSI-EDS data (*Jany et al.*, 2017; *Kotula et al.*, 2003; *Rossouw et al.*, 2016; *Rossouw et al.*, 2015; *Teng & Gauvin*, 2020). These algorithms aim to extract the underlying features from the available HSI-EDS data by reducing the dimensionality of the data, where high-dimensional pixel-wise datapoints are linearly projected onto a basis in a low-dimensional space (*Hotelling*, 1933; *Kotula et al.*, 2003; *Potapov & Lubk*, 2019; *Tipping & Bishop*, 1999). With these algorithms, phase masks are typically produced, which divide the dataset into regions belonging to the different components of the MSA models. Although able to perform without *a priori* assumptions, this approach contains inherent mathematical limitations (e.g., the restrictions of orthogonality and parsimony), which may lead to non-intuitive and non-interpretable results (*Kotula et al.*, 2003; *Stork & Keenan*, 2010). Clustering has been explored as an alternative approach. Clustering is an unsupervised technique that organizes entities into clusters or groups whose members bear similarities (*Funk et al.*, 2001; *Rui & Wunsch*, 2005; *Stork & Keenan*, 2010). Some clustering algo-

rithms, such as k-means and fuzzy clustering, are commonly applied for phase characterization (*Durdziński et al.*, 2015; *MacRae et al.*, 2007; *Vekemans et al.*, 2004; *Yan et al.*, 2006) (*Duan et al.*, 2016; *Parish*, 2019). Nevertheless, such algorithms have some intrinsic drawbacks. For example, k-means has problems analyzing data with varying sizes and densities. Whilst fuzzy clustering does allow pixels to belong to multiple clusters, the values of the pixel within each cluster are not probabilistic, providing little or no information on the confidence of the assignment.

In recent years, modern machine learning (ML) techniques have been successfully applied to analyzing electron microscopy datasets (*Ede*, 2021), including image denoising (*Antczak*, 2018; *Han et al.*, 2021; *Yoon et al.*, 2019), image classification (*Aguiar et al.*, 2019; *Vasudevan et al.*, 2018; *Yokoyama et al.*, 2020), and semantic segmentation (*Roberts et al.*, 2019; *Roels & Saeys*, 2019; *Urakubo et al.*, 2019; *Yu et al.*, 2020). However, most ML models are trained with human-labeled datasets in a supervised manner. To some extent, this makes the process still expensive and time-consuming, limiting the ability to unlock the potential of these methods. Self-supervised learning can avoid these problems by acquiring supervisory signals from the data itself (*Yann & Ishan*, 2021). In self-supervised learning, models are trained to capture the underlying patterns of the input data without relying on labels (*Yann & Ishan*, 2021). Thus, the combination of self-supervised or unsupervised algorithms, i.e., dimensionality reduction and clustering, can leverage the inherent structure in the HSI data to explore or identify physically sensible features (*Chen et al.*, 2021).

In this work, we introduce a self-supervised ML approach that automatically identifies unknown phases and unmixes the overlapped chemical signals for each potential phase with only one HSI-EDS dataset. This approach leverages a neural network autoencoder to extract underlying features of data through dimensionality reduction. A probabilistic Gaussian mixture model is used to identify inherent clusters, followed by factor analysis through non-negative factorization to distinguish chemical signals from the background. We name this new approach Spectral Interpretation using Gaussian Mixtures and Autoencoder (SIGMA). It is shown that SIGMA works on various HSI-EDS datasets with no need for user expertise in machine learning while bringing a significant improvement of accuracy and reliability. An overview of this approach is illustrated in Figure 1.

Here, we evaluate SIGMA using two HSI-EDS datasets, both motivated by the types of data typically encountered in studies of particulate matter air pollution. Such samples pose a particular challenge to interpretation using HSI-EDS as they comprise complex mixtures of unknown, overlapping phases, deposited in an uncontrolled manner on non-ideal, non-planar substrates (e.g., air filters or leaf substrates), and commonly contain grains that are smaller than the electron beam interaction volume. The two samples chosen are: (1) a synthetic mixture containing seven known minerals; (2) a sample representing a potential source of vehicular particulate matter. The synthetic mixture sample dataset

demonstrates the reliability and accuracy of this approach. SIGMA is further examined using the real-world particulate matter dataset, where the complex nature of the sample complicates the identification of the individual pollution particles. Additionally, SIGMA is built into a user-friendly Python code and can produce results within 30 min for a regular computer or even faster using graphic processing units (GPUs).

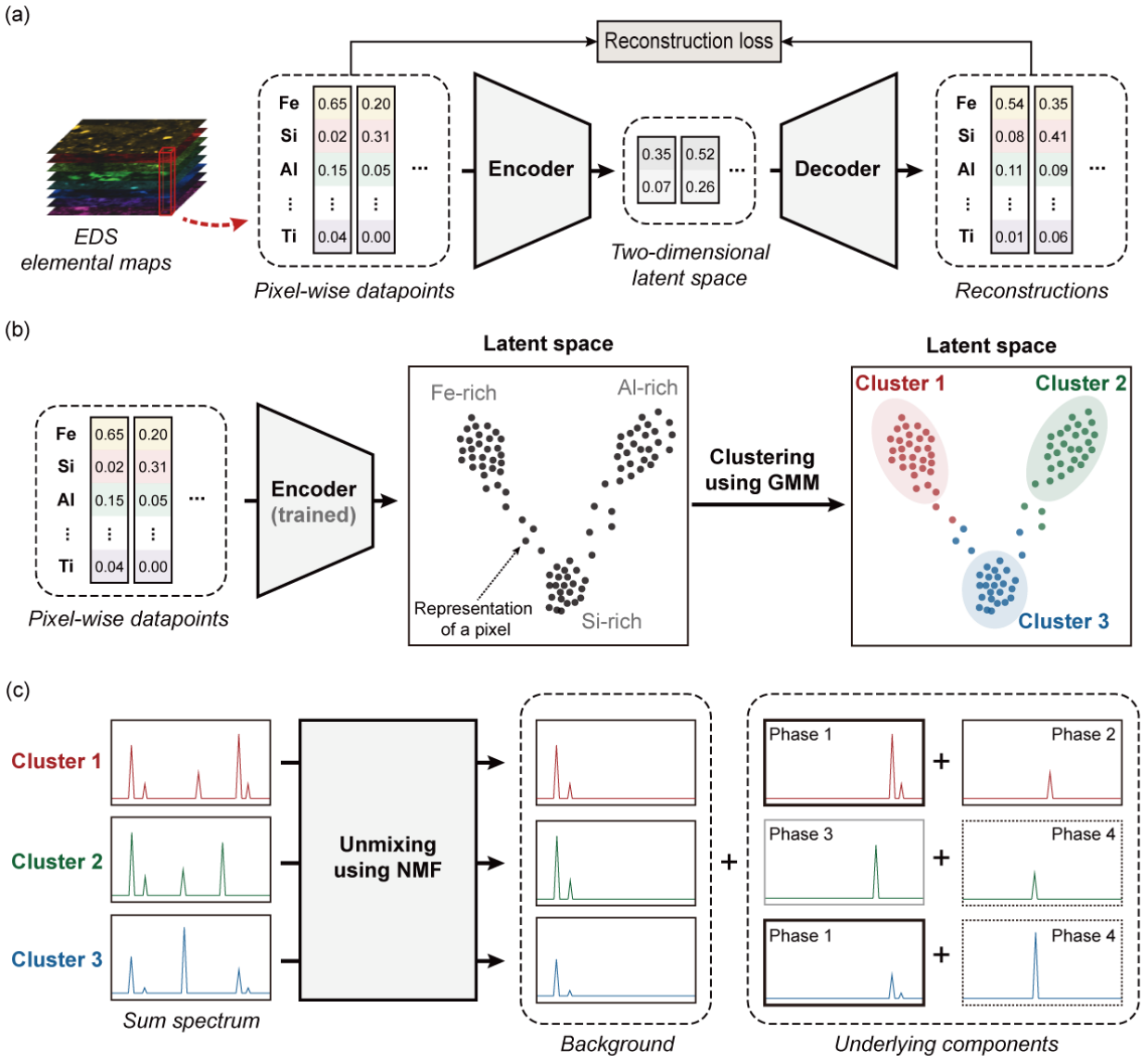

Figure 1. Workflow of SIGMA showing phase identification and signal unmixing on an HSI-EDS dataset. (**a**) A neural network autoencoder is trained to

learn good representations of elemental pixels in the 2D latent space. (**b**) The trained encoder is then used to transform high-dimensional elemental pixels into low-dimensional representations, followed by clustering using Gaussian mixture modeling (GMM) in the informative latent space. (**c**) Non-negative matrix factorization (NMF) is applied to unmix the single-phase spectra for all clusters. In such a way, the algorithm not only identifies the locations of all unknown phases but also isolates the background-subtracted EDS spectra of individual phases.

2 Background

Throughout the paper, scalars are represented by italics, e.g., k. Vectors and matrices are represented by boldface lowercase characters, e.g., $\mathbf{x}$, and boldface uppercase characters, e.g., $\mathbf{X}$, respectively.

2.1 Neural network autoencoder

Autoencoder is a neural network architecture that consists of two neural networks: an encoder and a decoder. The encoder $f_\phi(\mathbf{x})$ with parameters converts the input $\mathbf{x}$ to a low-dimensional representation $\mathbf{z}$, and the decoder $g_\theta(\mathbf{z})$ with parameters attempts to map the representation $\mathbf{z}$ back to a reconstruction of the initial input $\hat{\mathbf{x}}$. Upon training, autoencoder aims to minimize the error in reproducing the initial input $\mathbf{x}$, i.e., the reconstruction loss:

$$L(\mathbf{x},\ \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \left\|\mathbf{x} - g_\theta\left(f_\phi(\mathbf{x})\right)\right\|^2$$

The critical attribute of designing an autoencoder is through an information bottleneck (*Tishby & Zaslavsky*, 2015). The bottleneck forces the model to learn a compressed representation that contains the underlying information of the data. As a result, autoencoder is often applied to dimensionality reduction (*Hinton & Salakhutdinov*, 2006). Due to its non-linear characteristic, the autoencoder can learn representations that capture more complicated features than traditional methods, such as PCA, which only employs linear transformation on the data.

2.2 Gaussian mixture modeling

Gaussian mixture modeling (GMM) is an unsupervised probabilistic technique that fits clusters as a linear superposition of K Gaussian distributions (*Bishop & Nasrabadi*, 2006), which can be expressed as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k N\left(\mathbf{x}|\ _k,\ _k\right)$$

where w_k is the weighting coefficient, and $N\left(\mathbf{x}|\ _k,\ _k\right)$ denotes the kth Gaussian component of the mixture and is parametrised with the mean $_k$ and the covariance $_k$. Clustering through GMMs is achieved by applying the maximum

likelihood via expectation-maximisation (EM) algorithm (*Bishop & Nasrabadi*, 2006; *Dempster et al.*, 1977), where the models attempt to learn optimal solutions for parameters (i.e., $_k$, $_k$ and w_k for each Gaussian distribution) to model the empirical data distribution. In GMM clustering, datapoints are probabilistically assigned to clusters, therefore providing the confidence of the assignment, which makes the clustering process physically meaningful.

3 Materials and Methods

3.1 Datasets

The synthetic mixture sample is composed of seven mineral phases, including calcium carbonate ($CaCO_3$), orthoclase feldspar ($KAlSi_3O_8$), magnetite (Fe_3O_4), aluminum oxide (Al_2O_3), silicon oxide (SiO_2), titanium oxide (TiO_2), and zinc carbonate ($ZnCO_3$). All mineral phases were ground into particles less than ~50 m, followed by individual measurements of EDS spectra for validation. All minerals were physically mixed, forming a synthetic mixture sample, and deposited onto carbon tape mounted on a standard aluminum SEM stub. The dimensions of the acquired EDS dataset are 279 × 514-pixel × 1547-spectral-channel.

The particulate matter specimen was collected by scraping the inside of an exhaust pipe of a petrol-powered vehicle in Lahore, Pakistan using an A5 paper. More details about the specimen can be found in (*Sheikh et al.*, 2022). The dimensions of the raw EDS dataset are 738 × 672-pixel × 1595-spectral-channel.

Both specimens were carbon-coated before collecting the EDS data to prevent charging. Backscattered electron (BSE) images and EDS raw data were collected at an accelerating voltage of 15 keV using a Thermofisher Quanta-650F scanning electron microscope at the University of Cambridge, Department of Earth Sciences, equipped with two Bruker XFlash 6 EDS detectors.

3.2 Autoencoder architecture

For both synthetic and particulate matter datasets (9 and 8 elemental channels, respectively), the encoder block consists of three fully connected layers with 512, 256, and 128 neurons, respectively. Each layer is followed by a layer normalization (LayerNorm) layer (*Ba et al.*, 2016) and a Leaky Rectified Linear Unit (LeakyReLU) with a slope of -0.02 as activation function. LayerNorm is a technique that normalizes distributions of neuron outputs in the intermediate layers of a neural network; it can enhance the training speed of neural networks (*Ba et al.*, 2016). Different from ReLU, which gives zeros as outputs for negative inputs, LeakyReLU outputs a small linear component for each negative input (in this case, inputs are multiplied by 0.02 for negative values). This provides small positive gradients for negative outputs during training, avoiding the "dying ReLU" issue (*Lu et al.*, 2019). The decoder block uses the reversed structure of the encoder. The autoencoder was trained with Adam (*Kingma & Ba*, 2014) as optimizer function and squared L2 norm as loss function. Figure 2 shows the training history of the autoencoder used for the synthetic mixture dataset.

The loss values for the training (85% data), validation (15% data), and all data datasets converge progressively within 100 epochs. Note that the autoencoder architecture can vary according to the number of pixels and the number of the elemental channels of the dataset. The proposed architecture is suitable for datasets with 8–11 input elemental channels, which falls in the regime of typical mineralogical analyses.

3.3 Gaussian mixture modeling parameters

One big challenge for GMM clustering is to determine the number of clusters. We use the Bayesian information criterion (BIC) (*Bishop & Nasrabadi*, 2006) and the elbow method (*Wit et al.*, 2012) to quantitively determine the optimal number of clusters. BIC is a metric that measures the trade-off between the model complexity and the goodness of fit (i.e., maximum likelihood) to the datapoints, which is defined as:

$$\mathrm{BIC} = p \ln(N) - 2 \ln\left(\widehat{L}\right)$$

where p is the number of parameters in the GMM model, N is the number of datapoints for the GMM clustering, and $\widehat{L}$ is the mean likelihood for the GMM model. The elbow method is to locate the "elbow" of the BIC curve as the optimal number of clusters (K) based on the law of diminishing marginal returns (*Wit et al.*, 2012). Figure 3 shows the result of the elbow method, where the optimal number of clusters is 12, i.e., when K>12, the model fitting does not benefit from the increase of the number of clusters.

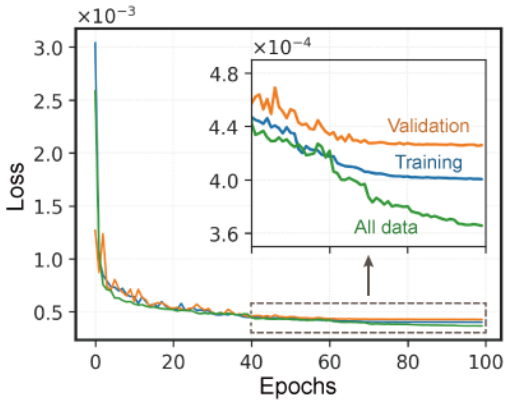

Figure 2. Training history for the autoencoder trained on the training data, validation data, and all data.

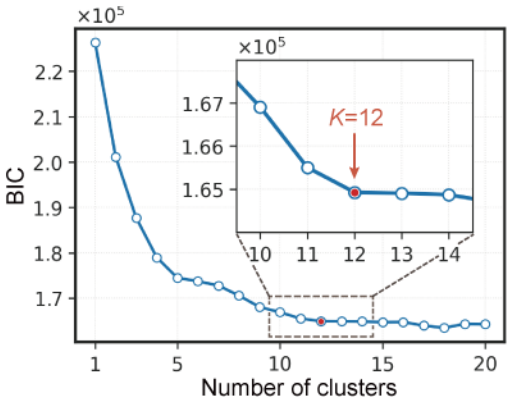

Figure 3. Bayesian information criterion (BIC) scores as a function of the number of clusters (K) of GMM, showing a preference for K=12, marked in red. Note that the data here is associated with the GMM clustering results in Figure 7.

4 Results and Discussion

4.1 Data pre-processing and normalization

The data pre-processing consists of three sequential steps: (optional) smoothing, z-score normalization, and softmax normalization. Figure 4 shows an example of the Fe signal intensity maps and the associated histograms after each pre-processing or normalization step.

Prior to the normalization steps, the synthetic mixture dataset is binned into the dimensions of 139 × 257-pixel × 1547-spectral-channel. Elemental intensity maps (i.e., X-ray lines of Al K , C K , C*a* K , Fe K , K K , O K , Si K , Ti K , and Zn L) are extracted, where the width of the energy windows is defined as the double full-width-at-half-maximum (FWHM) of individual elemental peaks with no background subtraction. This yields a datacube with the size of 139 × 257 × 9 for further processing. Each elemental map (with the size of 139 × 257 × 1) is then smoothed individually by applying a 3 × 3 mean filter, where each pixel is replaced by the average of pixel values in the surrounding 3 × 3 pixel area, as shown in Figure 4b.

Then, z-score normalization is separately applied to each elemental map (with the size of 139 × 257 × 1), converting the mean and the standard deviation of the intensity values into 0 and 1 in each elemental map, respectively, as shown in Figure 4c. Consequently, for each elemental intensity map, regions with intensity values above the average will become positive, while the other way around will become negative. With respect to a single 9-channel pixel (with the size of 1 × 1 × 9), the higher the positive value is, the more "above-average" the element composition is, in comparison with the same channel of other pixels. With z-score normalization, pixels in each elemental map incorporate the elemental

information across the entire measured area.

Next, each 9-channel pixel (regarded as a vector with the size of 1 × 1 × 9) is normalized to 0–1 interval using the softmax function. Softmax function (*Bishop & Nasrabadi*, 2006), or normalized exponential function, is a function that maps a feature vector of real values into a vector of probabilities that sum to one, which can be expressed as:

$$\mu_n = \frac{exp(\eta_n)}{\sum_j exp(\eta_j)}$$

where η_n represents the nth value in a feature vector , and μ_n represents the nth probability in a vector of probabilities . In the current case, as shown in Figure 4d, each 9-channel pixel vector (with the size of 1 × 1 × 9) will be transformed into a probability vector. Due to the characteristic of the exponential function, channels in a pixel with positive z-scores are emphasized, and those with negative z-scores are downplayed. Therefore, the values in a 9-channel pixel indicate the relative degrees of "above average" for individual elements. This can help the following machine learning model to extract underlying features from the dataset. Figure 5 displays elemental intensity maps of the synthetic mixture dataset after the sequential pre-processing and normalization steps.

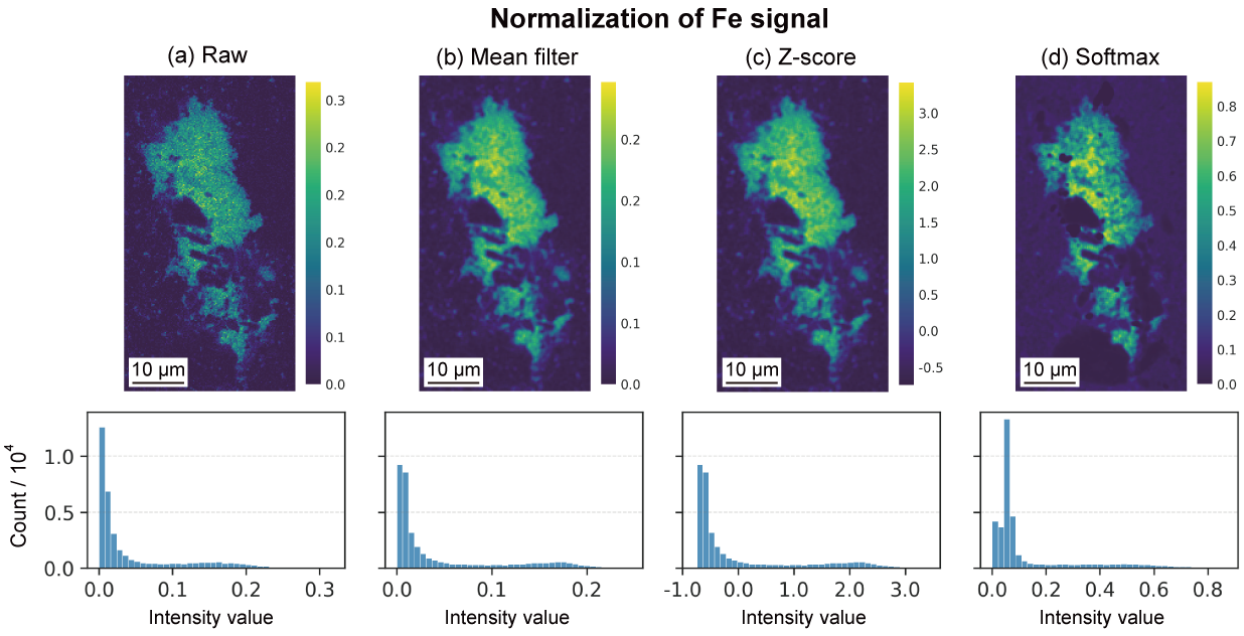

Figure 4. Elemental intensity maps and the associated histograms: (**a**) raw data; (**b**) smoothed data using a 3 × 3 mean filter; (**c**) data after smoothing and z-score normalization, and (**d**) data after smoothing, z-score, and softmax normalization.

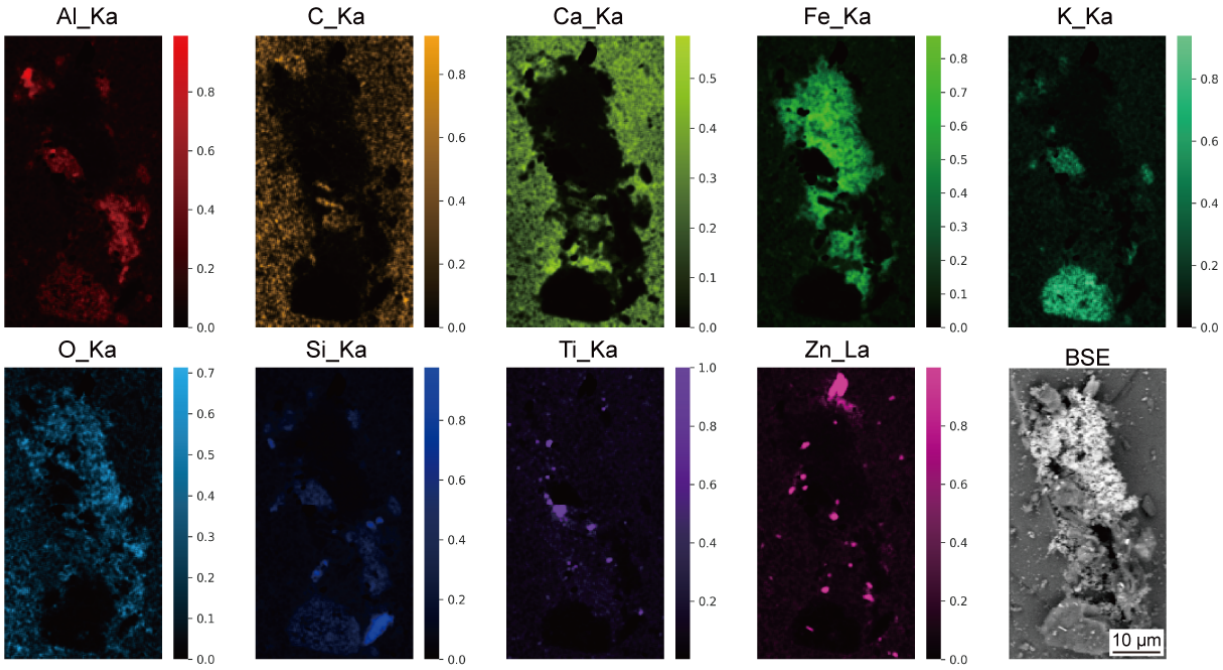

Figure 5. Normalized elemental intensity maps after the sequential pre-processing and normalization techniques, i.e., smoothing, using a 3×3 mean filter, and normalization using z-score and softmax. The associated backscattered electron (BSE) image of the same measured area.

4.2 Non-linear dimensionality reduction

We firstly use a neural network autoencoder to reduce the dimensionality of the nine-dimensional (9D) datapoints (9-elemental-channel pixels) before clustering. Although clustering directly in the 9D space is feasible, it might suffer from the problem of the curse of dimensionality (*Bellman et al.*, 1957; *Molchanov & Linsen*, 2018), i.e., datapoints in the high-dimensional space become sparse. In the current case, initial 9D pixels that belong to the same cluster may be still far from each other in the 9D space, limiting the performance of clustering algorithms. Some dimensionality reduction methods, such as PCA, are typically used to deal with this problem but usually come with the compromise of information loss and mixture of the underlying clusters during the process. Non-linear dimensionality reduction methods, such as autoencoder, can overcome these problems.

Figure 6 compares the ability of PCA and autoencoder to capture the underlying 2D structure of the synthetic mixture dataset (35,723 datapoints). In Figure 6a, datapoints are linearly projected onto the first two principal components with the highest variances, forming a distribution with a three-pointed-star shape.

Only three clusters are conceivable in the PCA-modelled latent space. On the other hand, the autoencoder (Figure 6b) splits datapoints into more clusters in the 2D latent space due to its capability to learn non-linear transformation. This may bring physical meaning to the latent space, which PCA lacks (discussed in the later section). Figure 6c shows the distribution of the pixel-wise datapoints in the autoencoder-modeled latent space, providing brief density information of the empirical distribution.

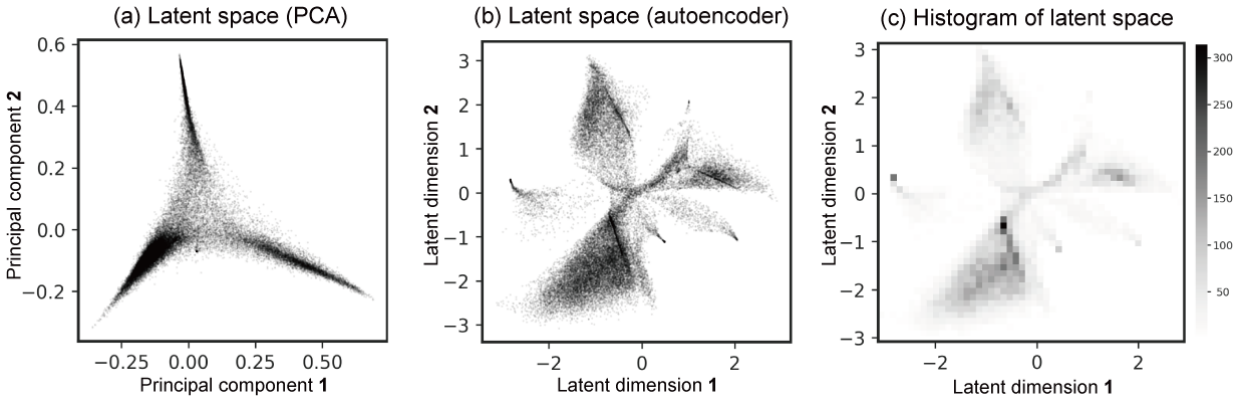

Figure 6. Two-dimensional visualizations of the synthetic mixture dataset. The latent space is modeled by (**a**) PCA by taking the first two principal components and (**b**) autoencoder, where each datapoint represents the associated pixel in the high-dimensional elemental intensity vector space; (**c**) the associated latent space histogram showing the datapoint distribution.

4.3 Clustering in two-dimensional latent space

We perform GMM clustering directly to the 2D representation of pixels in the latent space modeled by the autoencoder. In this process, chemically similar pixels are grouped into the same cluster. Figure 7 shows the clustering result for a GMM having K=12 components, where datapoints in different clusters are marked in different colors with 95% confidence ellipses superimposed. Each datapoint is assigned to the cluster for which the posterior probability $p(C_k|\mathbf{x})$ is the highest, i.e., given a datapoint $\mathbf{x}$, the probability that it belongs to the cluster C_k is the highest. The clustering result yields areas that point out compositional differences, which makes the latent space physically meaningful. For example, cluster #8, located in the middle of the latent space, contains a similar elemental signal to the averaged signals of all pixels, indicating no elemental fluctuation in this area. On the other hand, clusters with one or two enriched elemental signals tend to locate in the margin of the latent space, e.g., clusters #1 and #2 in the upper middle are Fe-rich, and cluster #7 on the left shows a strong Zn signal. Interestingly, the elemental signals of pixels increase from the center to the point within clusters, i.e., the composition of a pixel will smoothly change from the average to element-rich signals. Furthermore,

gradual transitions of elemental intensity among clusters are observed; the Al signal decreases as the cluster changes from cluster #3 to #4 to #5. Four unlabelled clusters that belong to background signals are circled with a dotted line.

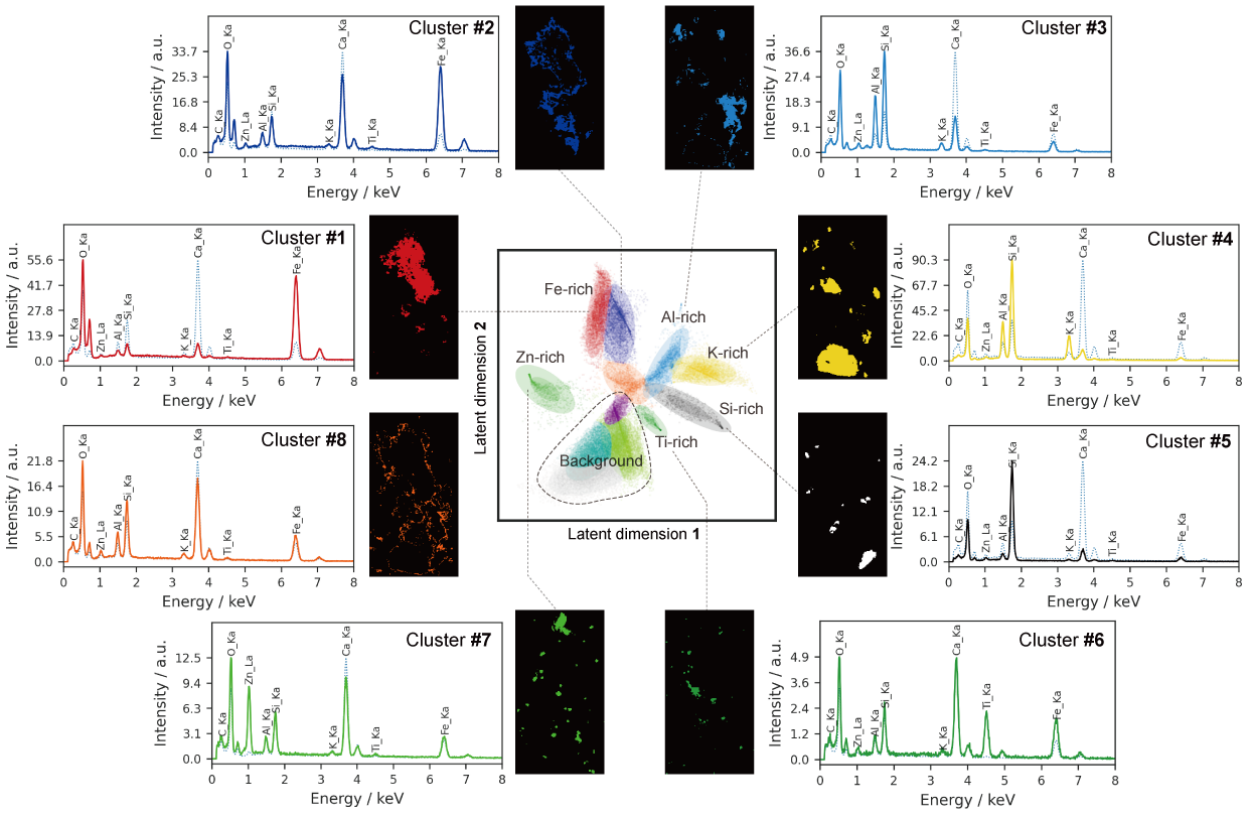

Figure 7. Visualization of latent space clustered using Gaussian mixture modeling. Each cluster is marked with a different color and overlapped with the associated 95% confidence ellipse. Locations and the sum EDS spectrum of the pixels in each cluster are illustrated. The blue dotted lines denote the average spectrum of all pixels in the synthetic mixture dataset. Note that the average spectrum is normalized to the scale of the sum spectra.

4.4 Unmixing overlapped EDS spectra

A key limitation of the GMM clustering result is that none of the cluster-spectra corresponds to the single-phase spectra measured separately. The detection of multiple-phase EDS signals can be explained by the unique surface morphology of the sample. In both synthetic mixture and particulate matter samples, mineral particles are spatially piled or stacked on top of each other. During EDS signal collection, these particles may contribute to the emission of X-rays

due to the electron-specimen interaction. As a result, each pixel may include signals from multiple phases and the background. Upon GMM clustering, the mixture of multiple-phase signals is observed in the sum spectra of pixels in each cluster (Figure 7). Thus, although having compositional signals, clusters still contain potential background and mixed-phase signals and fail to match any single-phase spectrum.

To obtain background-subtracted signals, we apply NMF to unmix the individual cluster-spectra. In this study, "unmixing" refers to distinguishing underlying EDS spectra of individual phases (called "components") from the superposed spectra that consist of a mixture of the contribution of each phase and determining the associated weights of each spectrum component (called "abundance"). The mixture of the spectra $\mathbf{x}_i$ is approximated using a linear mixing model (*Bioucas-Dias et al.*, 2012):

$$\mathbf{x}_i = \sum_{i=1}^{k} a_i \mathbf{s}_i + \mathbf{n}$$

where $\mathbf{s}_i$ is the underlying components of individual spectra, a_i is the abundance coefficients, and $\mathbf{n}$ is additive noise. Different from typical MSA approaches that analyze the pixel-wise dataset (*Benhalouche et al.*, 2019; *Kotula et al.*, 2003), NMF is applied here to the sum spectra of the GMM-clusters (called "cluster-spectra") on the synthetic mixture dataset, i.e., 1547×12 data matrix $\mathbf{X}$ that consists of 12 cluster-spectra with 1547 spectral-channels. The NMF for unmixing can be formulated as:

$$\mathbf{X} \approx \mathrm{SA}$$

where the 1547×12 matrix $\mathbf{S}$ is composed of 12 pseudo-spectra components (i.e., $\mathbf{s}_i$ in the linear mixing model), and the 12×12 matrix $\mathbf{A}$ contains the associated abundance coefficients (i.e., a_i in the linear mixing model). In this case, the NMF is applied without involving dimensionality reduction (i.e., the number of pseudo-spectra components is equal to the number of clusters) because no prior knowledge is provided. The optimal approximation of $\mathbf{X}$ is obtained through minimizing the Frobenius norm of the matrix difference. In addition, a regularization term (i.e., elementwise L1 norm) is applied to penalizes the model yielding a trivial solution (i.e., $\mathbf{S} = \mathbf{X}$ and $\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix) and facilitates more sparse solutions. Thus, the objective function for the unmixing NMF can be expressed as:

$$\min_{\mathbf{S},\ \mathbf{A}\geq 0} \| \mathbf{X} - \mathrm{SA} \|_F^2 + \mathrm{R}(\mathbf{S}) = \min_{\mathbf{S},\ \mathbf{A}\geq 0} \sum_{i,\ j} \left(\mathbf{X}_{\mathrm{ij}} - (\mathbf{SA})_{\mathrm{ij}} \right)^2 + \rho \sum_{i,r} \mathbf{S}_{\mathrm{ir}}$$

where ρ is a hyperparameter determining the impact of the regularization term. Figure 8 shows the unmixed 12 pseudo-spectra components. To examine the

accuracy of the unmixing performance, we compare these 12 pseudo-spectra with the real single-phase spectra that are measured separately. All seven phases are identified, and the associated spectra show an average cosine similarity of 83.0% to the experimental single-phase spectra, where the ground truth spectra are normalized to the same scales as the pseudo-spectra.

However, the unmixing process is not perfect. First, some phases are not properly unmixed, e.g., TiO_2 and $ZnCO_3$ in components #10 and #12, respectively. This may result from the relatively small amount of these phases in the dataset. In the GMM clustering step, pixels with similar elemental signals are grouped into the same cluster. Clusters that include pixels from minor phases would have lower signal intensity in the sum spectrum, e.g., the Ti-rich cluster #6 only contains 455 pixels yielding the Ti peak with height ~2.4 a.u., whereas the Fe-rich cluster #1 contains 3,936 pixels yielding the Fe peak with height ~41.7 a.u. (as shown in Figure 7). This leads to imbalanced signal intensity scales among different cluster-spectra, e.g., the intensities of peaks in the Fe-rich cluster are much higher than in the Ti-rich cluster. Consequently, cluster-spectra with major phases (higher intensity scales) tend to acquire better approximation upon optimization with the criterion of the Frobenius norm. In contrast, cluster-spectra with minor phases (lower intensity scales) may yield relatively inaccurate unmixed pseudo-spectra, or even be overlooked by the algorithm. Second, some components may have little or no physical meaning, i.e., showing compositions with unrealistic intensity ratios and/or a combination of elemental peaks. These components do not correspond to any ground truth phase and may represent the general background and/or noise introduced by instrument artifacts. For instance, component #7 (containing only the potassium peak) does not fit any measured phase and is regarded as part of the signal from $KAlSi_3O_8$. Also, components #4, showing a strong oxygen signal, may be interpreted as the component that capture instrumental artifacts. Third, some components are repetitive. For example, components #9 and #11 are similar to component #3 (Fe_3O_4) but have extra peaks.

These problems can be mitigated by analyzing the abundance coefficients (a_i) and the intensity of peaks in the pseudo-spectra. According to the linear mixing model, each cluster can be approximated by a linear combination of underlying spectra weighted by abundance coefficients. Figure 9 shows the analysis of the abundance coefficients for each component, indicating the importance of the contribution of each component. As shown in Figure 9a, cluster #1 can be approximated using components #3 and #4 with an abundance coefficient of 13.9 and 5.9, respectively; component #4 is responsible for the oxygen signal in component #3. Therefore, cluster #1 most likely is Fe_3O_4. In most cases, abundance coefficients are sparse, i.e., only one or two components are dominant, as shown in Figure 9b, c, and d. As a result, most physically meaningless components with low abundance coefficients may be intrinsically excluded when drawing inferences. Similar abundance coefficient analyses can be conducted for all clusters, producing a phase map for the synthetic mixture dataset (Figure 10).

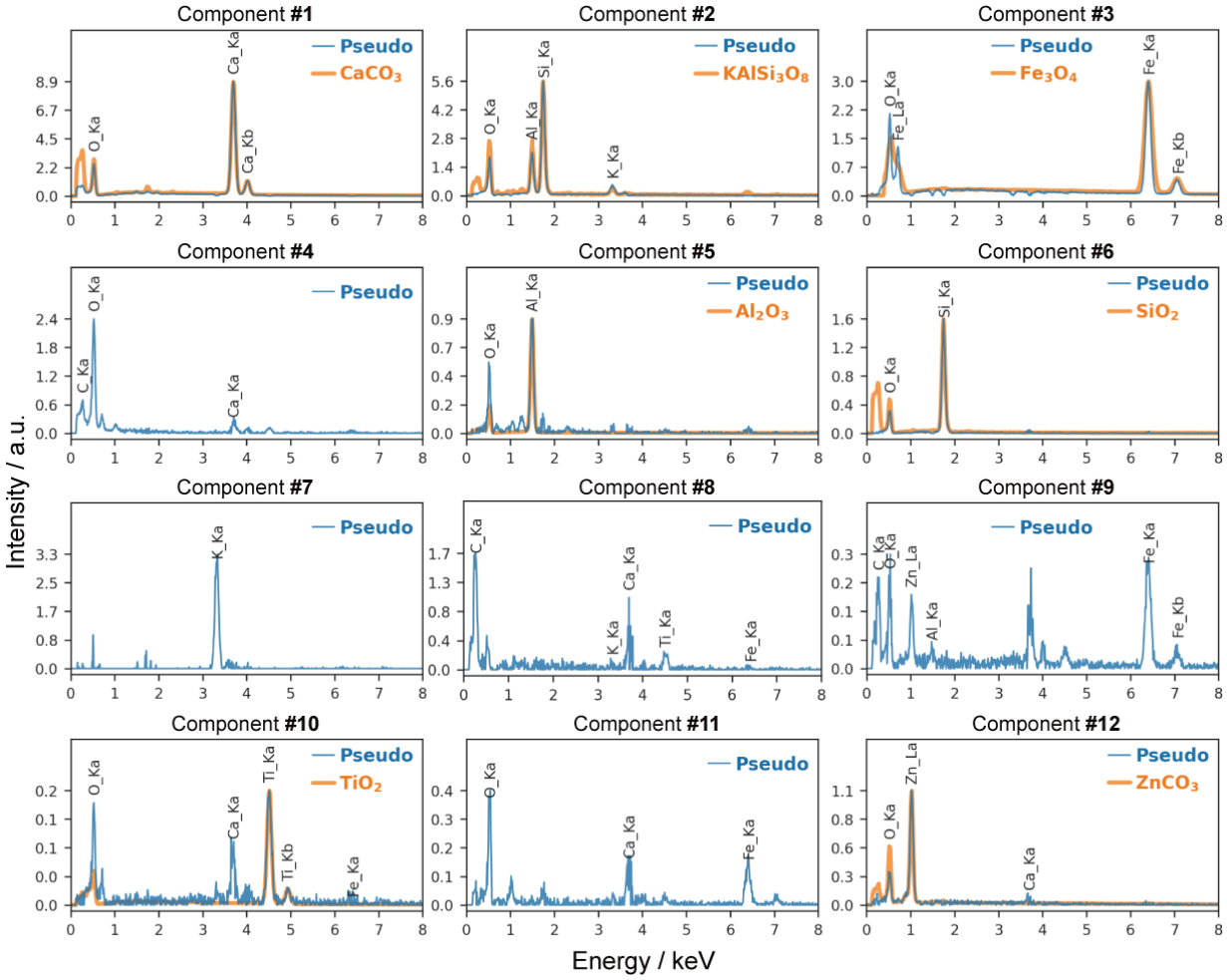

Figure 8. NMF components showing the underlying pseudo-spectra. Certain pseudo-spectra components are in excellent agreement (i.e., average cosine similarity = 83.0%) with ground truth single-phase spectra measured from the individual mineral particles before the mixture. Note that the real single-phase spectra are marked in orange and normalized to the same scales as the associated pseudo-spectra.

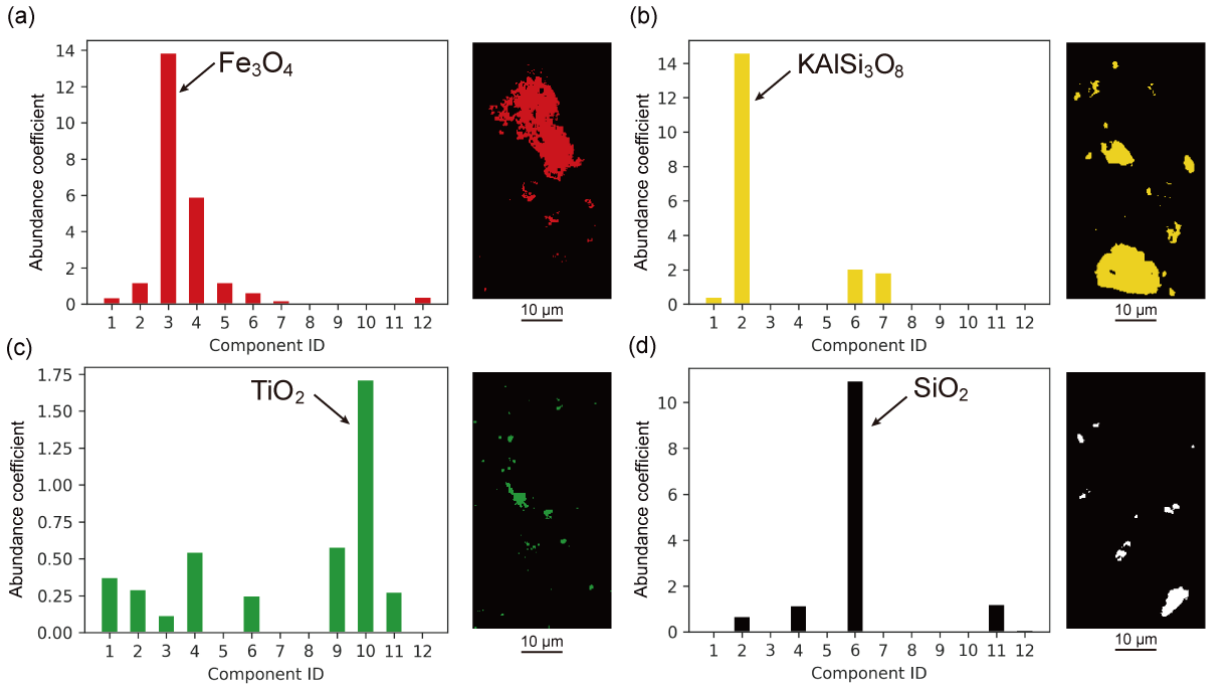

Figure 9. Bar charts of abundance coefficients and pixel distributions showing the importance of NMF components for each cluster. The underlying single-phase spectrum of each cluster of (**a**) Fe_3O_4, (**b**) $KAlSi_3O_8$, (**c**) TiO_2, and (**d**) SiO_2 can be identified according to the abundance coefficients.

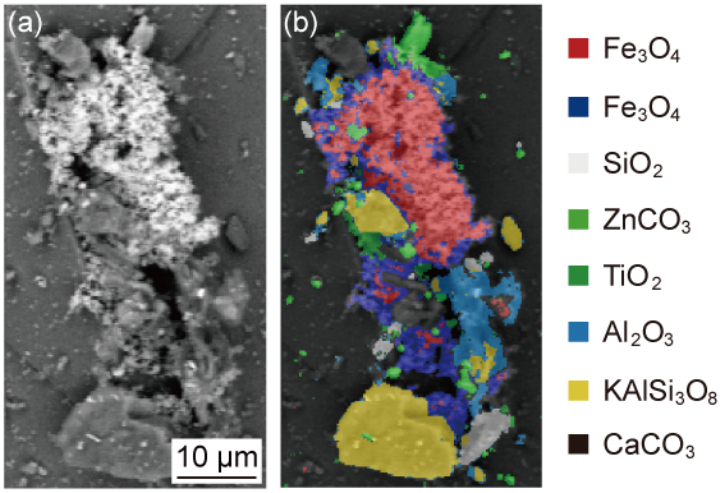

Figure 10. Backscattered electron (BSE) image of the synthetic mixture dataset and the corresponding phase map according to the NMF unmixing anal-

ysis. Note that red and dark blue represent the same phase of Fe_3O_4 but are unmixed from different clusters.

4.5 Testing SIGMA on exhaust pipe residue particulate matter dataset

We evaluate the performance of SIGMA on the exhaust pipe residue particulate matter dataset, where particles with various compositions and sizes are distributed on the substrate. Exposure to particles containing heavy metals, particularly, Fe-bearing ultrafine particles can have serious health implications. Inhalation of Fe-rich nanoparticles is a major health risk for cardiovascular diseases (Dusseldorp et al., 1995; Maher et al., 2020) and has been found to enter the human brain through olfactory transport (Maher et al., 2016). The toxicity of Fe-bearing ultra-fine particles is linked to their size, composition, and distribution. Thus, it is critical to identify and quantify the abundant presence of Fe-bearing ultrafine particles in urban microenvironments. In a previous study (*Sheikh et al.*, 2022), the task to identify these particles was manually conducted by individually analyzing backscattered electron (BSE) images and their associated EDS elemental maps, which is an inefficient and time-consuming process.

Here, SIGMA offers huge potential for automated identification of potential Fe-bearing particles with background-subtracted compositional information (Figure 11). After pre-processing and normalizing (the same procedure in the synthetic mixture dataset), the dataset is built into elemental intensity maps with the dimensions of 396 × 336-pixel × 8-spectral-channel (i.e., X-ray lines of O K , Fe K , Mg K , Ca K , Al K , C K , Si K , and S K). Then, an autoencoder is trained to learn the 2D representation of pixels. Figure 11a shows the autoencoder-modeled 2D latent space and the result of GMM clustering (K=13), where datapoints that belong to different clusters are marked with different colors. Again, the clusters yield physically meaningful areas, indicating compositional information for pixels. The distribution of pixels forms a pointed-star shape, where the center refers to the averaged signals, and the arms represent certain element-rich clusters.

In this specimen, our main goal was to identify Fe-bearing ultrafine particles; therefore, we primarily focus on the Fe-rich phase (green cluster observed in the top right of the latent space). Figure 11b and c show the spatial distribution of Fe-rich pixels and the associated backscattered electron (BSE) image. The size distribution (Figure 11d) is obtained by analyzing the spatial distribution of the identified Fe-bearing particles in Figure 11b. Prior to the unmixing step, although containing the Fe K peak well above the average, the sum spectrum (Figure 11d) appears to include overlapped signals from the background. After NMF unmixing, the background-subtracted Fe-oxide spectrum is successfully identified through abundance coefficient analysis (Figure 11f). We can see that SIGMA is capable of not only identifying potential Fe-bearing particles but also unmixing and isolating its chemical signal from the matrix in an automated manner.

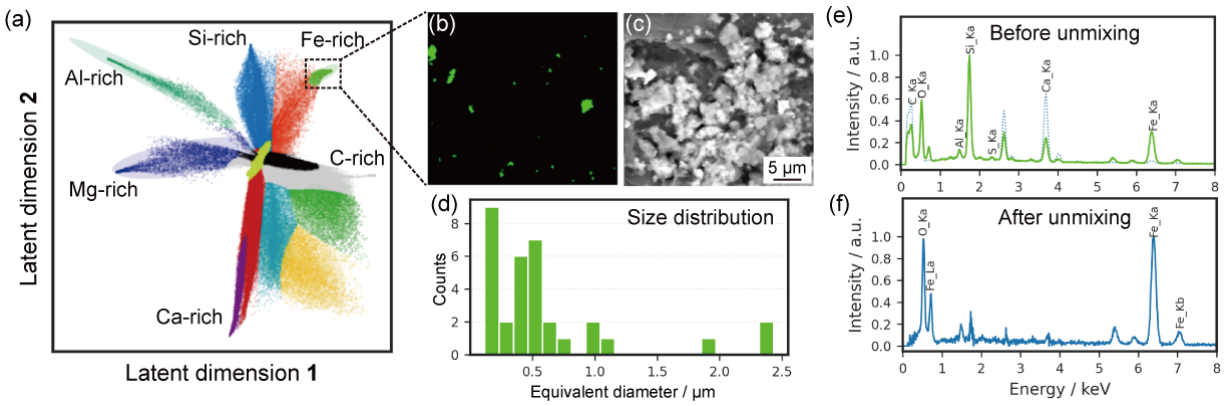

Figure 11. Phase identification and elemental signal unmixing on particulate matter dataset using SIGMA. (**a**) 2D latent space modeled by autoencoder showing the GMM clustering result, where clusters are marked with different colors and overlapped with the associated 95% confidence ellipses; (**b**) pixel distribution in the Fe-rich cluster; (**c**) associated backscattered electron (BSE) image; (**d**) size distribution of the Fe-rich particles, where the equivalent diameter is defined as the diameter of the circle that has the same area of the region. Normalized sum spectrum of the Fe-rich cluster (**e**) before NMF unmixing (overlaid with the average spectrum of the blue dotted line) and (**f**) after NMF unmixing.

5 Conclusions

We have developed a self-supervised approach for automated phase identification and hyperspectral unmixing with only one hyperspectral image—energy dispersive X-ray spectroscopy (HSI-EDS) dataset. Specifically, we apply non-linear dimensionality reduction to the HSI dataset using a neural network autoencoder and analyze the underlying structure of the data using Gaussian mixture modeling (GMM) clustering. Non-negative matrix factorization (NMF) is employed cluster-by-cluster to isolate the background-subtracted EDS signals from the matrix. We evaluate this approach with two HSI-EDS datasets. For the known synthetic mixture dataset, all seven major phases are identified and verified by the individually measured EDS spectra, revealing the accuracy (i.e., average cosine similarity = 83.0%) and robustness of our technique. For the particulate matter dataset, the performance of this approach is further demonstrated by distinguishing potential Fe-bearing particles from several unknown chemical phases with different particle sizes. Furthermore, the proposed approach can be applied to more general HSI datasets, such as electron energy loss spectroscopy (EELS), scanning tunneling microscopy (STM), and time-of-flight secondary ion mass spectrometry (ToF-SIMS), providing a robust and reliable analysis in a fully automated manner.

Acknowledgments

P.-Y. Tung and R. J. Harrison acknowledge funding by the Electron and X-

ray microscopy Community for structural and chemical Imaging Techniques for Earth materials (EXCITE) (award number - G106564). F. Nabiei was supported by the Swiss National Science Foundation (SNSF) postdoctoral mobility fellowships P2ELP2_184386. P.-Y. Tung thanks Prof. Paul Midgley for the fruitful discussion. P.-Y. Tung and H. A. Sheikh are grateful for the synthetic mixture sample preparation from Dr Giulio Lampronti. H. A. Sheikh appreciates the Cambridge Trust for PhD funding.

Data Availability Statement

We wrapped the entire SIGMA as a Python module and have built it into a user-friendly notebook with GUI (available at Zenodo: https://doi.org/10.5281/zenodo.6468991).

References

Adam, E., O. Mutanga, and D. Rugege (2010), Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review, *Wetlands Ecology and Management*, *18*(3), 281-296.Afromowitz, M. A., J. B. Callis, D. M. Heimbach, L. A. DeSoto, and M. K. Norton (1988), Multispectral imaging of burn wounds: a new clinical instrument for evaluating burn depth, *IEEE Transactions on Biomedical Engineering*, *35*(10), 842-850.Aguiar, J., M. L. Gong, R. Unocic, T. Tasdizen, and B. Miller (2019), Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning, *Science advances*, *5*(10), eaaw1949.Antczak, K. (2018), Deep recurrent neural networks for ECG signal denoising, *arXiv preprint arXiv:1807.11551*.Ba, J. L., J. R. Kiros, and G. E. Hinton (2016), Layer normalization, *arXiv preprint arXiv:1607.06450*.Bellman, R., R. Corporation, and K. M. R. Collection (1957), *Dynamic Programming*, Princeton University Press.Benhalouche, F. Z., M. S. Karoui, and Y. Deville (2019), An NMF-Based Approach for Hyperspectral Unmixing Using a New Multiplicative-tuning Linear Mixing Model to Address Spectral Variability, paper presented at 2019 27th European Signal Processing Conference (EUSIPCO), 2-6 Sept. 2019.Bioucas-Dias, J. M., A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot (2012), Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, *IEEE journal of selected topics in applied earth observations and remote sensing*, *5*(2), 354-379.Bishop, C. M., and N. M. Nasrabadi (2006), *Pattern recognition and machine learning*, Springer.Blackburn, G. A. (2006), Hyperspectral remote sensing of plant pigments, *Journal of Experimental Botany*, *58*(4), 855-867.Bosman, M., M. Watanabe, D. T. L. Alexander, and V. J. Keast (2006), Mapping chemical and bonding information using multivariate analysis of electron energy-loss spectrum images, *Ultramicroscopy*, *106*(11), 1024-1032.Carrasco, O., R. Gomez, A. Chainani, and W. Roper (2003), *Hyperspectral imaging applied to medical diagnoses and food safety*, SPIE.Chen, Z., et al. (2021), Machine learning on neutron and x-ray scattering and spectroscopies, *Chemical Physics Reviews*, *2*(3), 031301.Dempster, A. P., N. M. Laird, and

D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.Duan, X., F. Yang, E. Antono, W. Yang, P. Pianetta, S. Ermon, A. Mehta, and Y. Liu (2016), Unsupervised Data Mining in nanoscale X-ray Spectro-Microscopic Study of NdFeB Magnet, *Scientific Reports*, *6*(1), 34406.Durdziński, P. T., C. F. Dunant, M. B. Haha, and K. L. Scrivener (2015), A new quantification method based on SEM-EDS to assess fly ash composition and study the reaction of its individual components in hydrating cement paste, *Cement and Concrete Research*, *73*, 111-122.Ede, J. M. (2021), Deep learning in electron microscopy, *Machine Learning: Science and Technology*, *2*(1), 011004.Feng, Y.-Z., and D.-W. Sun (2012), Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review, *Critical Reviews in Food Science and Nutrition*, *52*(11), 1039-1058.Funk, C. C., J. Theiler, D. A. Roberts, and C. C. Borel (2001), Clustering to improve matched filter detection of weak gas plumes in hyperspectral thermal imagery, *IEEE Transactions on Geoscience and Remote Sensing*, *39*(7), 1410-1420.Gendrin, C., Y. Roggo, and C. Collet (2008), Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, *Journal of Pharmaceutical and Biomedical Analysis*, *48*(3), 533-553.Goldstein, J. I., D. E. Newbury, J. R. Michael, N. W. Ritchie, J. H. J. Scott, and D. C. Joy (2017), *Scanning electron microscopy and X-ray microanalysis*, Springer.Govender, M., K. Chetty, and H. Bulcock (2007), A review of hyperspectral remote sensing and its application in vegetation and water resource studies, *Water SA*, *33*(2), 145-151.Gowen, A. A., C. P. O'Donnell, P. J. Cullen, G. Downey, and J. M. Frias (2007), Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends in Food Science & Technology*, *18*(12), 590-598.Han, Y., et al. (2021), Deep learning STEM-EDX tomography of nanocrystals, *Nature Machine Intelligence*, *3*(3), 267-274.Hinton, G. E., and R. R. Salakhutdinov (2006), Reducing the Dimensionality of Data with Neural Networks, *Science*, *313*(5786), 504-507.Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of educational psychology*, *24*(6), 417.Jany, B. R., A. Janas, and F. Krok (2017), Retrieving the Quantitative Chemical Information at Nanoscale from Scanning Electron Microscope Energy Dispersive X-ray Measurements by Machine Learning, *Nano Letters*, *17*(11), 6520-6525.Kannan, R., A. V. Ievlev, N. Laanait, M. A. Ziatdinov, R. K. Vasudevan, S. Jesse, and S. V. Kalinin (2018), Deep data analysis via physically constrained linear unmixing: universal framework, domain examples, and a community-wide platform, *Advanced Structural and Chemical Imaging*, *4*(1), 6.Kingma, D. P., and J. Ba (2014), Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.Kotula, P. G., M. R. Keenan, and J. R. Michael (2003), Automated Analysis of SEM X-Ray Spectral Images: A Powerful New Microanalysis Tool, *Microscopy and Microanalysis*, *9*(1), 1-17.Lu, L., Y. Shin, Y. Su, and G. E. Karniadakis (2019), Dying relu and initialization: Theory and numerical examples, *arXiv preprint arXiv:1903.06733*.MacRae, C., N. Wilson, A. Torpy, U. Rossek, and M. Rohde (2007), The Holistic Approach to Data Collection and Hyperspectral Analysis, *Microscopy and Microanalysis*, *13*(S02), 1358-1359.Malinowski, E. R., and D.

G. Howery (1980), *Factor analysis in chemistry*, Wiley New York.Molchanov, V. F., and L. Linsen (2018), Overcoming the Curse of Dimensionality When Clustering Multivariate Volume Data, paper presented at VISIGRAPP.Parish, C. M. (2019), Fuzzy Clustering to Merge EDS and EBSD Datasets with Crystallographic Ambiguity, *Microscopy and Microanalysis*, *25*(S2), 134-135.Potapov, P., and A. Lubk (2019), Optimal principal component analysis of STEM XEDS spectrum images, *Advanced Structural and Chemical Imaging*, *5*(1), 4.Roberts, G., S. Y. Haile, R. Sainju, D. J. Edwards, B. Hutchinson, and Y. Zhu (2019), Deep learning for semantic segmentation of defects in advanced STEM images of steels, *Scientific reports*, *9*(1), 1-12.Roels, J., and Y. Saeys (2019), Cost-efficient segmentation of electron microscopy images using active learning, *arXiv preprint arXiv:1911.05548*.Rossouw, D., B. R. Knappett, A. E. H. Wheatley, and P. A. Midgley (2016), A New Method for Determining the Composition of Core–Shell Nanoparticles via Dual-EDX+EELS Spectrum Imaging, *Particle & Particle Systems Characterization*, *33*(10), 749-755.Rossouw, D., P. Burdet, F. de la Peña, C. Ducati, B. R. Knappett, A. E. H. Wheatley, and P. A. Midgley (2015), Multicomponent Signal Unmixing from Nanoheterostructures: Overcoming the Traditional Challenges of Nanoscale X-ray Analysis via Machine Learning, *Nano Letters*, *15*(4), 2716-2720.Rui, X., and D. Wunsch (2005), Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, *16*(3), 645-678.Sheikh, H. A., B. A. Maher, V. Karloukovski, G. I. Lampronti, and R. J. Harrison (2022), Biomagnetic Characterization of Air Pollution Particulates in Lahore, Pakistan, *Geochemistry, Geophysics, Geosystems*, *23*(2), e2021GC010293.Stork, C. L., and M. R. Keenan (2010), Advantages of Clustering in the Phase Classification of Hyperspectral Materials Images, *Microscopy and Microanalysis*, *16*(6), 810-820.Teng, C., and R. Gauvin (2020), Multivariate Statistical Analysis on a SEM/EDS Phase Map of Rare Earth Minerals, *Scanning*, *2020*, 2134516.Tipping, M. E., and C. M. Bishop (1999), Probabilistic principal component analysis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611-622.Tishby, N., and N. Zaslavsky (2015), Deep learning and the information bottleneck principle, paper presented at 2015 ieee information theory workshop (itw), IEEE.Urakubo, H., T. Bullmann, Y. Kubota, S. Oba, and S. Ishii (2019), UNI-EM: an environment for deep neural network-based automated segmentation of neuronal electron microscopic images, *Scientific reports*, *9*(1), 1-9.Vasudevan, R. K., N. Laanait, E. M. Ferragut, K. Wang, D. B. Geohegan, K. Xiao, M. Ziatdinov, S. Jesse, O. Dyck, and S. V. Kalinin (2018), Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images, *npj Computational Materials*, *4*(1), 1-9.Vekemans, B., L. Vincze, F. E. Brenker, and F. Adams (2004), Processing of three-dimensional microscopic X-ray fluorescence data, *Journal of Analytical Atomic Spectrometry*, *19*(10), 1302-1308.Wit, E., E. v. d. Heuvel, and J.-W. Romeijn (2012), 'All models are wrong...': an introduction to model uncertainty, *Statistica Neerlandica*, *66*(3), 217-236.Yan, B., T. R. McJunkin, D. L. Stoner, and J. R. Scott (2006), Validation of fuzzy logic method for automated mass spectral classification for mineral imaging, *Applied Surface Science*, *253*(4), 2011-2017.Yann, L., and M. Ishan (2021), Self-supervised learning: The dark matter

of intelligence, edited, Meta AI Blog.Yokoyama, Y., T. Terada, K. Shimizu, K. Nishikawa, D. Kozai, A. Shimada, A. Mizoguchi, Y. Fujiyoshi, and K. Tani (2020), Development of a deep learning-based method to identify "good" regions of a cryo-electron microscopy grid, *Biophysical Reviews*, *12*(2), 349-354.Yoon, D., H. S. Lim, K. Jung, T. Y. Kim, and S. Lee (2019), Deep learning-based electrocardiogram signal noise detection and screening model, *Healthcare informatics research*, *25*(3), 201-211.Yu, Z. X., S. C. Wei, J. W. Zhang, B. Wang, Y. J. Wang, Y. Liang, and H. L. Tian (2020), High-throughput, algorithmic determination of pore parameters from electron microscopy, *Computational Materials Science*, *171*, 109216.Zaefferer, S., and G. Habler (2017), Scanning electron microscopy and electron backscatter diffraction, in *Mineral reaction kinetics: Microstructures, textures, chemical and isotopic signatures*, edited by W. Heinrich and R. Abart, p. 0, European Mineralogical Union and Mineralogical Society of Great Britain and Ireland.