Geochemistry, Geophysics, Geosystems®

AGU ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE
10.1029/2022GC010530

SIGMA: Spectral Interpretation Using Gaussian Mixtures and Autoencoder

Po-Yen Tung[1,2], Hassan A. Sheikh[1], Matthew Ball[1], Farhang Nabiei[1,3], and 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

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 data set with an average spectrum similarity 83% to the ground truth
- The approach demonstrates its ability to generalize by identifying unknown minerals and their background-subtracted chemical signals

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P.-Y. Tung,
pyt21@cam.ac.uk

Citation:

Tung, P.-Y., Sheikh, H. A., Ball, M., Nabiei, F., & Harrison, R. J. (2023). SIGMA: Spectral interpretation using Gaussian mixtures and autoencoder. *Geochemistry, Geophysics, Geosystems*, *24*, e2022GC010530. https://doi.org/10.1029/2022GC010530

Received 17 MAY 2022
Accepted 9 JAN 2023

Author Contributions:

Conceptualization: Po-Yen Tung, Matthew Ball, Farhang Nabiei
Data curation: Po-Yen Tung, Hassan A. Sheikh
Formal analysis: Po-Yen Tung
Funding acquisition: Richard J. Harrison
Investigation: Po-Yen Tung
Methodology: Po-Yen Tung, Farhang Nabiei
Project Administration: Richard J. Harrison
Software: Po-Yen Tung, Farhang Nabiei
Supervision: Richard J. Harrison

Abstract Identification of unknown micro- and nano-sized mineral phases is commonly achieved by analyzing chemical maps generated from hyperspectral imaging data sets, 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 data points 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 data sets: 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 flexible and adaptable 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 data set, 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 micro-analysis 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 data set. 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 data sets. 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

Validation: Po-Yen Tung, Hassan A. Sheikh, Richard J. Harrison
Visualization: Po-Yen Tung
Writing – original draft: Po-Yen Tung
Writing – review & editing: Po-Yen Tung, Hassan A. Sheikh, Farhang Nabiei, Richard J. Harrison

exploration of the HSI-EDS data (Jany et al., 2017; Kotula et al., 2003; Rossouw et al., 2015, 2016; 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 data points 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 data set 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 centroid-based clustering algorithms, such as *k*-means and fuzzy clustering, are commonly applied for phase characterization (Duan et al., 2016; Durdziński et al., 2015; MacRae et al., 2007; Martineau et al., 2019; Parish, 2019; Vekemans et al., 2004; Yan et al., 2006). Nevertheless, such algorithms have some intrinsic drawbacks for HSI data. For example, *k*-means has problems analyzing data with varying sizes and densities. Whilst fuzzy clustering does allow pixels to belong to multiple clusters and yield probabilistic interpretations, the values of the pixel within each cluster are usually based on Euclidean distance to the centroids, which is not always appropriate for measuring the similarity between data points (e.g., when the shapes of clusters are non-flat manifolds) (Wang et al., 2002). On the other hand, clustering algorithms that use non-Euclidean metrics can measure the non-Euclidean relationships between data points and may offer better performance. For instance, a density-based clustering algorithm (i.e., HDBSCAN, McInnes et al., 2017) has recently been applied to tackle the issue of data with varying densities (Blanco-Portals et al., 2022; Li et al., 2019). Other types of clustering techniques, such as distribution-based algorithms, have attracted less attention for analyzing HSI-EDS data so far.

In recent years, modern machine learning (ML) techniques have been successfully applied to analyzing electron microscopy data sets (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, the generalization of the workflows to different material systems, data types, and measurement conditions may require sizable modification. One of the common aims for electron microscopic studies is to extract the underlying phase or structural information from electron microscopic data without a priori knowledge. To this end, unsupervised or self-supervised learning emerges as a plausible solution. Self-supervised learning is a type of algorithm that acquires 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). Autoencoder is a neural network that can be used for self-supervised tasks. Autoencoders can learn low-dimensional representations efficiently by copying its input to its output (Hinton & Salakhutdinov, 2006). With autoencoders, the dimensionality of the data can be reduced without losing essential features. Despite being widely employed in other types of HSI data (Lin et al., 2013; C. Tao et al., 2015; X. Tao et al., 2022) and some electron microscopic data (Ede, 2020; McAuliffe et al., 2020), autoencoders have had limited applications in HSI-EDS data.

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 data set. 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 data sets with no need for user expertise in machine learning while bringing a significant improvement of accuracy and reliability.

Here, we evaluate SIGMA using two HSI-EDS data sets, 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: (a) a synthetic mixture

containing seven known minerals and (b) a sample representing a potential source of vehicular particulate matter. The synthetic mixture sample data set demonstrates the reliability and accuracy of this approach. SIGMA is further examined using the real-world particulate matter data set, 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).

2. Materials and Methods

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

2.1. Data Sets

Before introducing the two individual data sets, we discuss two inherent features of the samples. First, the two raw EDS data sets have relatively low average counts per pixel (a) to demonstrate the ability of SIGMA to extract meaningful information from low-quality data and (b) to suit the need for quick EDS analyses (or analyses over large areas), where the precision of measurements is subjected to statistical error. Second, the two samples demonstrated in this study are of the same type, that is, mineral particles are deposited on a substrate. In this case, particles are spatially stacked on top of each other, forming a unique morphology. As a result, upon data acquisition, the EDS spectrum collected in each pixel most likely includes emissions of X-rays from multiple mineral particles and the background, as the electron-specimen interaction volume is larger than the volume of the individual particles. The desire to unmix these overlapped signals further is the motivation for using an additional NMF step subsequent to the clustering step (as discussed later).

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 μm, followed by individual measurements of single-phase EDS spectra for validation. Note that the single-phase EDS spectra (denoted as ground truth spectra later) were measured from separate samples where only pure minerals exist. Then, 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 data set are 279 × 514-pixel × 1,547-spectral-channel, and the average counts per pixel is 29.94.

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 data set are 738 × 672-pixel × 1,595-spectral-channel, and the average counts per pixel is 35.08.

Both specimens were carbon-coated before collecting the EDS data to prevent charging. Backscattered electron (BSE) images 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. EDS raw data were measured using two Bruker XFlash 6|30 EDS detectors installed in the same SEM.

2.2. Data Pre-Processing and Normalization

The data pre-processing consists of three sequential steps: (optional) smoothing, z-score normalization, and softmax normalization. Figure 1 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 data set is binned into the dimensions of 139 × 257-pixel × 1,547-spectral-channel. Elemental intensity maps (i.e., X-ray lines of Al Kα, C Kα, Ca Kα, Fe Kα, K Kα, O Kα, Si Kα, Ti Kα, and Zn Lα) 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

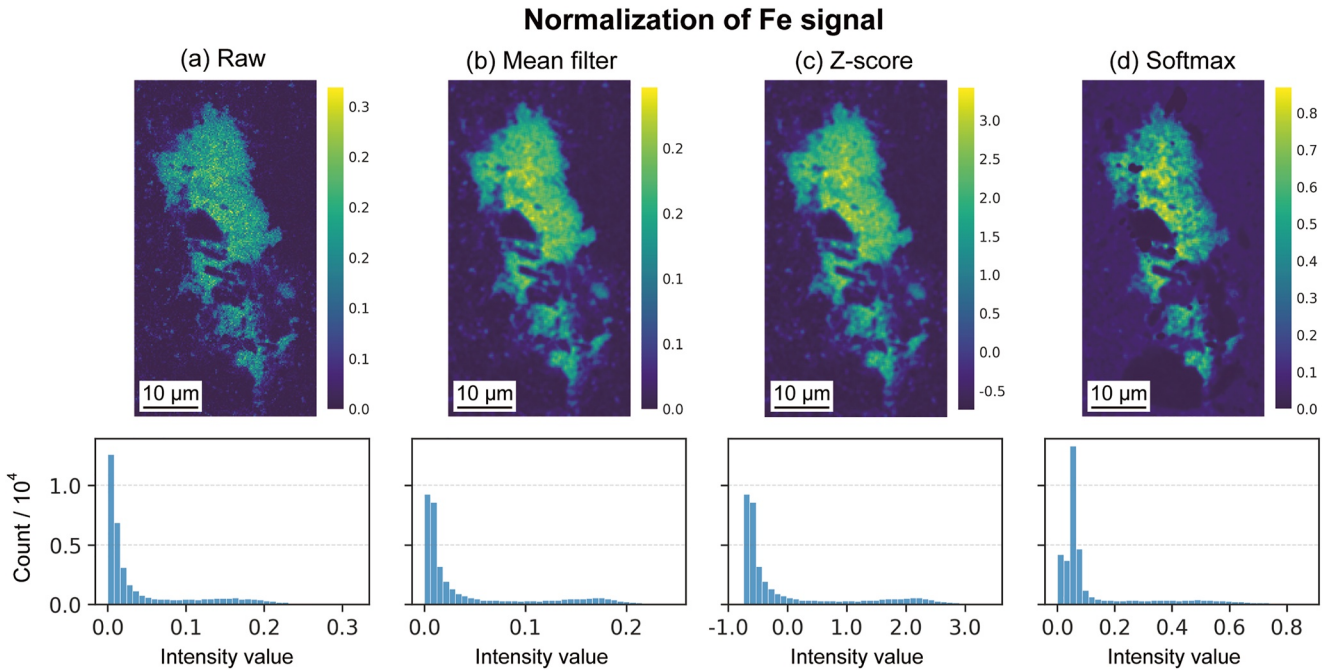

Figure 1. Elemental intensity maps and the associated histograms: (a) raw data; (b) smoothed data using a 3 × 3 mean filter; (c) data after smoothing and *z*-score normalization, and (d) data after smoothing, *z*-score, and softmax normalization.

the average of pixel values in the surrounding 3 × 3 pixel area, as shown in Figure 1b. Note that the key elements here are determined by manually identifying the presence of peaks in the sum spectrum of all pixels.

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 1c. Consequently, for each elemental intensity map, regions with intensity values above the average will become positive, while intensity values lower than average 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 $\boldsymbol{\eta}$ into a vector of probabilities $\boldsymbol{\mu}$ 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 $\boldsymbol{\eta}$, and μ_n represents the nth probability in a vector of probabilities $\boldsymbol{\mu}$. In the current case, as shown in Figure 1d, 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 data set. Figure 2 displays elemental intensity maps of the synthetic mixture data set after the sequential pre-processing and normalization steps.

2.3. Overview of SIGMA Workflow

An overview of the SIGMA workflow is illustrated in Figure 3. SIGMA aims to identify the phases and their single-phase spectra from HSI-EDS data sets. To this end, SIGMA is designed with three primary

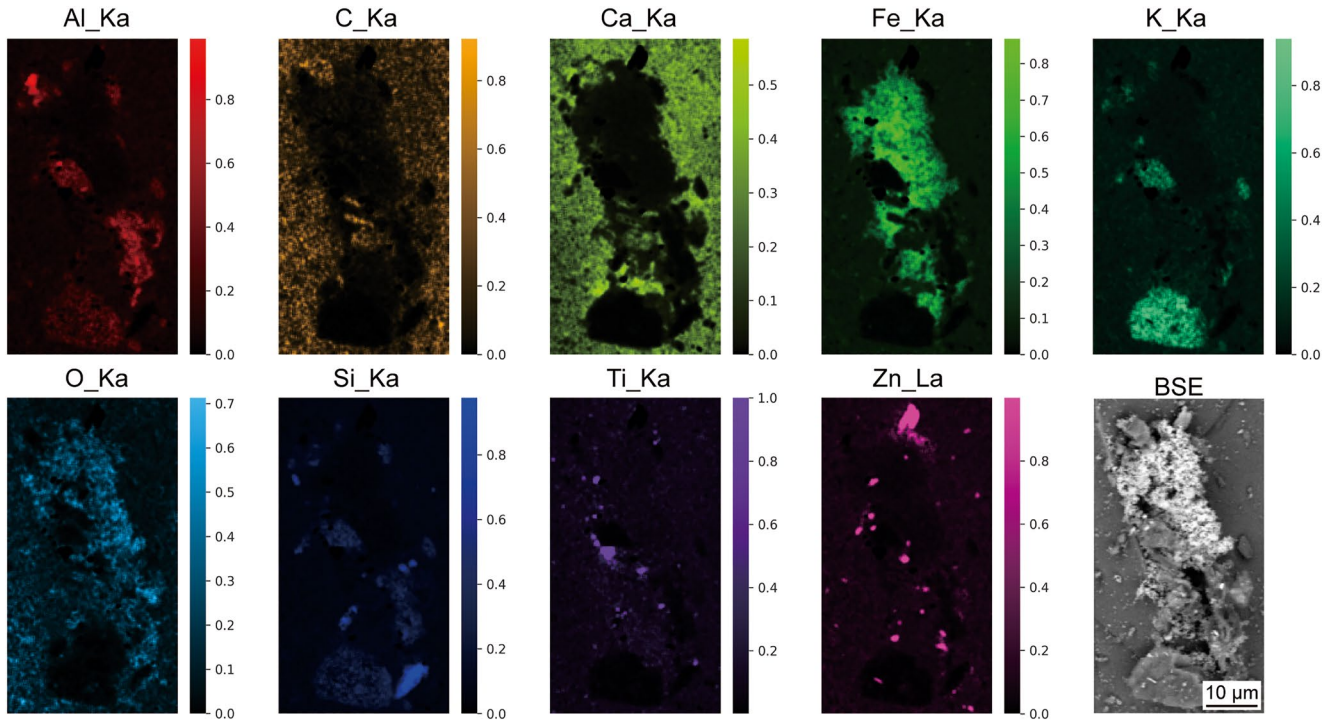

Figure 2. Normalized elemental intensity maps after the sequential pre-processing and normalization techniques, that is, smoothing, using a 3×3 mean filter, and normalization using z-score and softmax. The associated backscattered electron (BSE) image of the same measured area.

steps: dimensionality reduction, clustering, and unmixing. First, a neural network autoencoder is trained to reduce the dimensionality of the input (i.e., the pre-processed and normalized pixels) but also to keep the relationships in the high-dimensional space. Then, high-dimensional elemental pixels are transformed into two-dimensional (2D) latent representations using the trained encoder (Figure 3a). In the second step, clustering is performed on the 2D representations using Gaussian mixture modeling (GMM) to distinguish chemically different clusters in the latent space. As a result, chemically similar pixels are grouped into several clusters (Figure 3b). After summing the EDS spectra of pixels within each cluster, we obtain a sum spectrum for each cluster (denoted as "cluster-spectrum" later). Consequently, the initial pixel-wise HSI-EDS data set (with the dimension of 139 × 257-pixel × 1,547-spectral-channel) can be simplified to several cluster-spectra (12-cluster × 1,547-spectral-channel for the current case, as discussed later in Section 3.3). In the third step, NMF is applied to unmix the single-phase spectra from the cluster-spectra (Figure 3c). In such a workflow, the algorithm not only identifies the locations of all unknown phases but also isolates the background-subtracted EDS spectra of individual phases.

2.4. 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 $\boldsymbol{\phi}$ converts the input $\mathbf{x}$ to a low-dimensional representation $\boldsymbol{z}$, and the decoder $g_\theta(\mathbf{z})$ with parameters $\boldsymbol{\theta}$ attempts to map the representation $\boldsymbol{z}$ back to a reconstruction of the initial input $\hat{\mathbf{x}}$. Upon training, autoencoder aims to minimize the error in reproducing the initial input $\mathbf{x}$, that is, the reconstruction loss:

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

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.

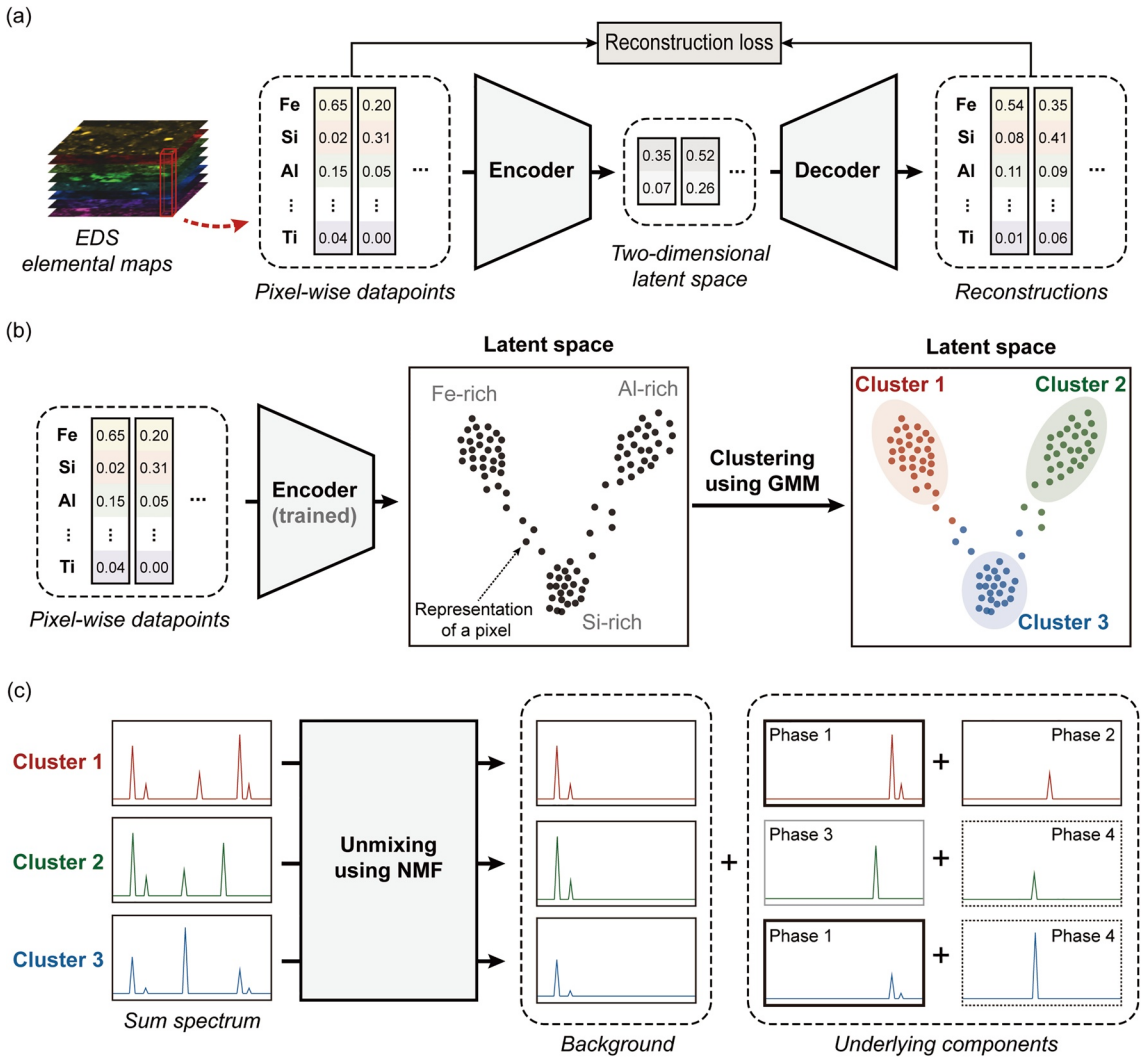

Figure 3. Workflow of SIGMA showing phase identification and signal unmixing on an HSI-EDS data set. (a) A neural network autoencoder is trained to learn good representations of elemental pixels in the 2D latent space. (b) The trained encoder is then used to transform high-dimensional elemental pixels into low-dimensional representations, followed by clustering using Gaussian mixture modeling in the informative latent space. (c) Non-negative matrix factorization is applied to unmix the single-phase spectra for all clusters.

For both synthetic and particulate matter data sets (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., 2020). 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 4a shows the training history of the autoencoder used for the synthetic mixture data set. The loss values for the training (85% data), validation (15% data), and all data sets 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 data set. Here, we used cross-validation for hyperparameter selection to determine recommended values that are appropriate for the vast majority of cases likely to be encountered by typical users. The proposed architecture is suitable for data sets with 8–11 input elemental channels, which falls in the regime of typical mineralogical analyses.

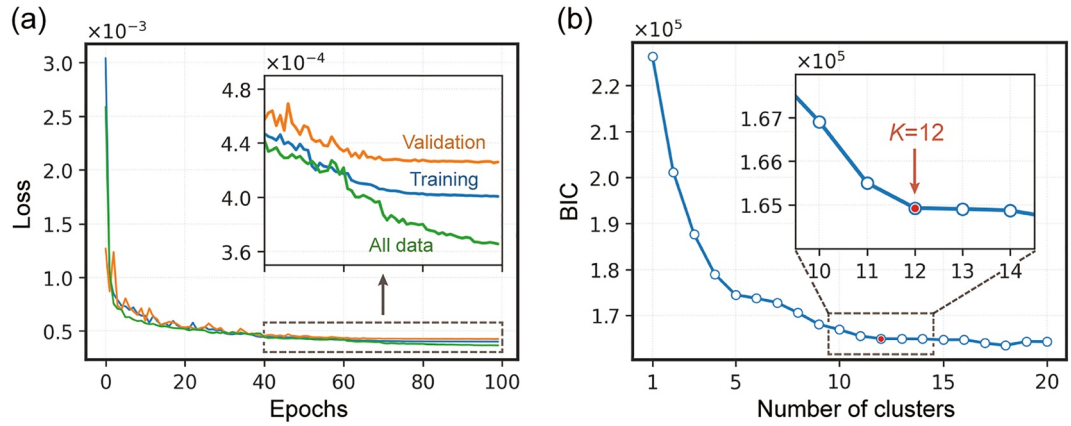

Figure 4. (a) Training history for the autoencoder trained on the training data, validation data, and all data. (b) Bayesian information criterion (BIC) scores as a function of the number of clusters (K) of GMM, showing a preference for $K = 12$, marked in red. Note that the data here is associated with the clustering results in Figure 6.

2.5. 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:

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

where w_k is the weighting coefficient, and $N\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ denotes the kth Gaussian component of the mixture and is parametrized with the mean $\boldsymbol{\mu}_k$ and the covariance $\boldsymbol{\Sigma}_k$. Clustering through GMMs is achieved by applying the maximum likelihood via expectation-maximization (EM) algorithm (Bishop & Nasrabadi, 2006; Dempster et al., 1977), where the models attempt to learn optimal solutions for parameters (i.e., $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and w_k for each Gaussian distribution) to model the empirical data distribution. In clustering using GMM, data points are probabilistically assigned to clusters, therefore providing the confidence of the assignment, which makes the clustering process physically meaningful.

One big challenge for clustering using GMM 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 data points, which is defined as:

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

where p is the number of parameters in the GMM model, N is the number of data points for the clustering using GMM, and $\hat{L}$ is the mean likelihood for the GMM model. Note that p is directly determined by the number of clusters. 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 4b shows the result of the elbow method, where the optimal number of clusters is 12, that is, when $K > 12$, the model fitting does not benefit from the increase of the number of clusters.

2.6. Matrix Factorization

Matrix factorization is a simple, non-parametric method that has been widely applied to discover underlying latent features of the data (He et al., 2005). In this study, PCA and NMF, two well-known matrix factorization techniques, are employed for the tasks of dimensionality reduction and factor analysis. PCA is used as a baseline method to evaluate the performance of the autoencoder for dimensionality reduction. NMF is included in the current workflow for unmixing the signals of sum EDS spectra processed by GMM.

2.6.1. Principal Component Analysis

Assuming directions with the largest variances having important features, principal component analysis (PCA) computes an orthogonal basis to re-express the given data set (Jolliffe, 2002; Potapov & Lubk, 2019; Shlens, 2014). The computed orthonormal basis vectors are referred to as principal components. One common way to figure out principal components is to apply singular value decomposition (SVD) to the data matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$$

where the $n \times m$ matrix $\mathbf{X}$ is composed of n elemental channels and m pixels, the columns of the $n \times m$ matrix $\mathbf{U}$ and the $m \times m$ matrix $\mathbf{V}$ are orthonormal, and the $m \times m$ matrix $\mathbf{\Sigma}$ is a diagonal matrix with positive values. For PCA, the columns of $\mathbf{U}$ represent the principal components, and rows of $\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$ are "scores" describing the contributions of the corresponding principal components to the data set.

By assuming that the most important features only remain in the principal components with the first i largest variances, one can conduct dimensionality reduction by dropping some of the principal components with lower variances (i.e., less informative principal components). In this study, although the heuristic interpretation of the Scree plot (Figure S1 in Supporting Information S1) suggests retaining the first three principal components with the three largest variances ($i = 3$), we opt to retain the first two principal components ($i = 2$), that is, the dimension of data points is reduced from nine to two, for direct comparison with the performance of dimensionality reduction using autoencoder.

2.6.2. Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) (Lee & Seung, 1999) decomposes a non-negative data matrix $\mathbf{X}$ into a product of non-negative matrices $\mathbf{S}$ and $\mathbf{A}$:

$$\mathbf{X} \approx \mathbf{S}\mathbf{A}$$

where the data matrix $\mathbf{X}$ consists of n pixels and m elemental channels (i.e., all pixels are included in the data matrix) in a typical MSA analysis using NMF (Kotula et al., 2003), the columns of the $n \times k$ matrix $\mathbf{S}$ represent components or latent factors, and the $k \times m$ matrix $\mathbf{A}$ contains the scores.

Here, instead of analyzing the pixel-wise data set, NMF is applied to unmixing the sum spectra of the GMM-clusters (called "cluster-spectra") on the synthetic mixture data set, that is, $1{,}547 \times 12$ data matrix $\mathbf{X}$ that consists of 12 cluster-spectra with 1,547 spectral-channels, the $1{,}547 \times 12$ matrix $\mathbf{S}$ is composed of 12 pseudo-spectra components, and the 12×12 matrix $\mathbf{A}$ contains the associated abundance coefficients. It should be noted that the NMF here 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 penalize 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:

$$\min_{\mathbf{S},\mathbf{A}\geq 0} \| \mathbf{X} - \mathbf{S}\mathbf{A} \|_F^2 + \rho \mathrm{R}(\mathbf{S}) = \min_{\mathbf{S},\mathbf{A}\geq 0} \sum_{i,j} (\mathbf{X}_{ij} - (\mathbf{S}\mathbf{A})_{ij})_2 + \rho \sum_{i,r} \mathbf{S}_{ir}$$

where ρ is a hyperparameter determining the impact of the regularization term.

2.7. Software and Package

The entire workflow of SIGMA was built in Python 3.7. Hyperspy 1.6.5 (https://doi.org/10.5281/zenodo.5608741) was employed to read and pre-process the HSI-EDS data sets. *Scikit-learn* 1.0.2 (Pedregosa et al., 2011) was used for PCA, NMF, and GMM. The autoencoder was implemented using *Pytorch* 1.10.0 (Paszke et al., 2019).

3. Results and Discussion

3.1. Non-Linear Dimensionality Reduction

We first use a neural network autoencoder to reduce the dimensionality of the nine-dimensional (9D) data points (9-elemental-channel pixels) before clustering. Although clustering directly in the 9D space is feasible, it might

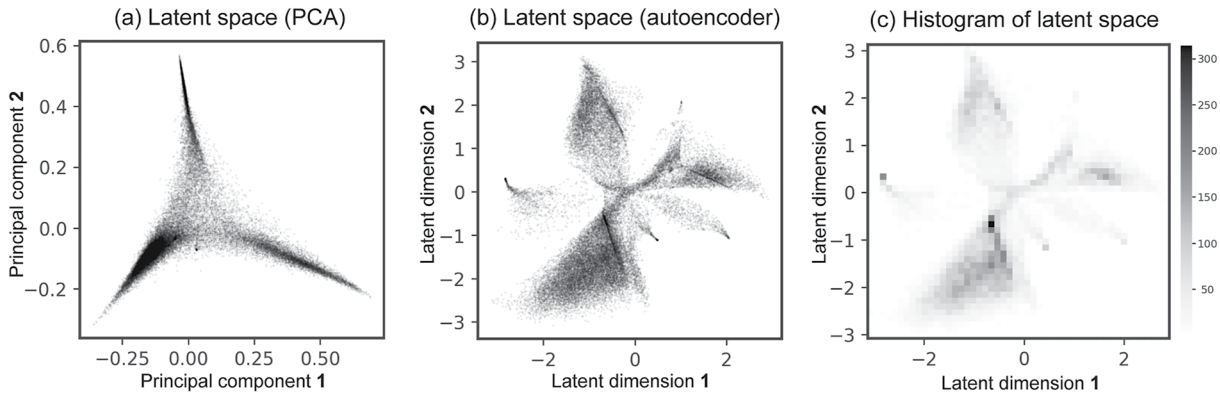

Figure 5. Two-dimensional visualizations of the synthetic mixture data set. The latent space is modeled by (a) PCA by taking the first two principal components and (b) autoencoder, where each data point represents the associated pixel in the high-dimensional elemental intensity vector space; (c) the associated latent space histogram showing the data point distribution.

suffer from the problem of the curse of dimensionality (Bellman et al., 1957; Molchanov & Linsen, 2018), that is, data points 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. Reducing the dimensionality of the data can facilitate better clustering results. The aim of dimensionality reduction is to reduce redundancy and figure out the underlying structure of the data (Sumithra & Surendran, 2015). The reduced representations should keep relations as much in the high dimensional space with minimal loss of information. Some linear 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. This assumes that the data can be represented as a linear combination of a smaller set of intrinsic components; it is not applicable if there exist non-linear relationships among high-dimensional data. On the other hand, non-linear dimensionality reduction methods, such as autoencoder, can overcome these problems.

Figure 5 compares the ability of PCA and autoencoder to capture the underlying 2D structure of the synthetic mixture data set (35,723 data points). In Figure 5a, data points 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-modeled latent space. On the other hand, the autoencoder (Figure 5b) splits data points 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 5c shows the distribution of the pixel-wise data points in the autoencoder-modeled latent space, providing brief density information of the empirical distribution.

3.2. 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 6 shows the clustering result for a GMM having $K = 12$ components, where data points in different clusters are marked in different colors with 95% confidence ellipses superimposed. Each data point is assigned to the cluster for which the posterior probability $p(C_k|\mathbf{x})$ is the highest, that is, given a data point $\mathbf{x}$, 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, for example, 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, that is, 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. Note that only clusters (and their sum spectra) that may contain meaningful mineral phases are labeled and presented; four unlabeled clusters that belong to background signals are circled with a dotted line.

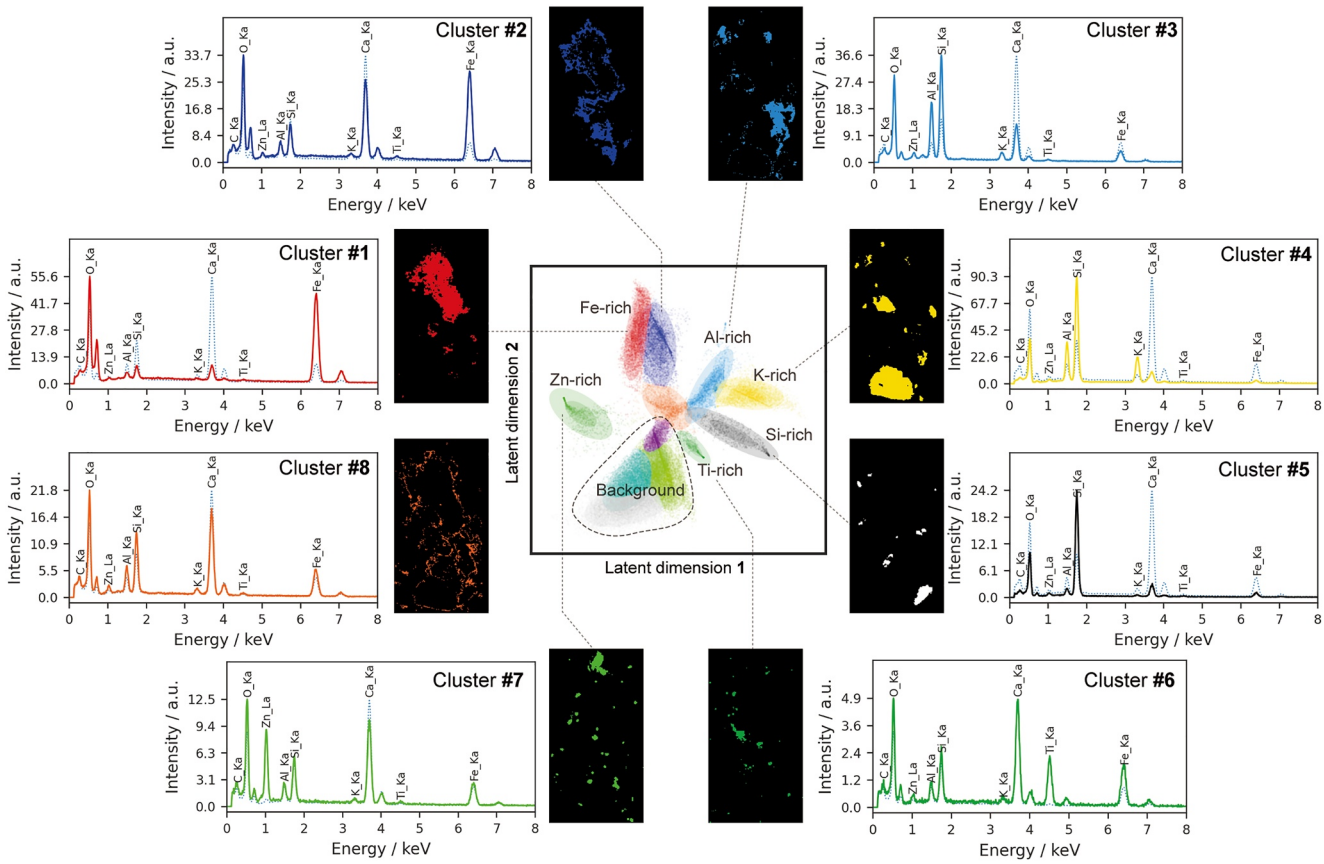

Figure 6. Visualization of latent space clustered using Gaussian mixture modeling. Each cluster is marked with a different color and overlapped with the associated 95% confidence ellipse. Locations and the sum EDS spectrum of the pixels in each cluster are illustrated. The blue dotted lines denote the average spectrum of all pixels in the synthetic mixture data set. Note that the average spectrum is normalized to the scale of the sum spectra.

3.3. 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 6). 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. To examine the accuracy of the unmixing performance, we compare these 12 pseudo-spectra Different from typical MSA approaches that analyze the pixel-wise data set (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

10.1029/2022GC010530

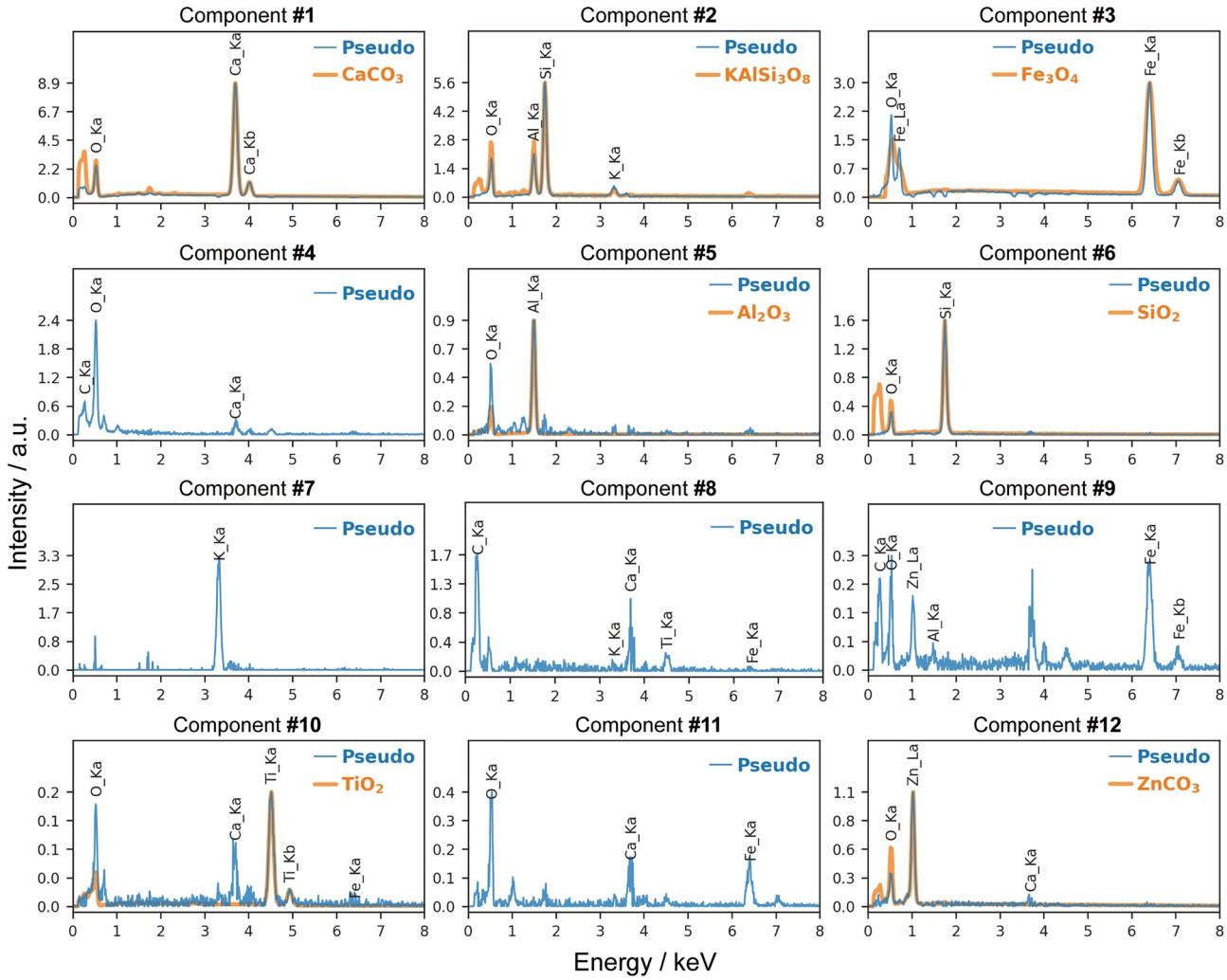

Figure 7. NMF components showing the underlying pseudo-spectra. Certain pseudo-spectra components are in excellent agreement (i.e., average cosine similarity = 83.0%) with ground truth single-phase spectra measured from the individual mineral particles before the mixture. Note that the real single-phase spectra are marked in orange and normalized to the same scales as the associated pseudo-spectra.

data set, that is, 1,547 × 12 data matrix $\mathbf{X}$ that consists of 12 cluster-spectra with 1,547 spectral-channels (see Methods for more details). This unmixing procedure is valid because the dimensionality reduction and clustering steps effectively simplify the input of NMF from a 1,547 × 35,723 data matrix (with noisy spectra) down to a 1,547 × 12 data matrix, each column of which is the sum of spectra from chemically similar pixels and, therefore, significantly reduces statistical unambiguity of spectra (i.e., becomes significantly less noisy). In this case, pixels in a GMM cluster are considered the same type of spatial mixture of mineral particles; for example, all pixels in cluster #2 are assumed to consist of Fe_3O_4 and $CaCO_3$. Thus, only 12 types of mixtures of mineral particles are assumed to exist (as $K = 12$ is specified for GMM) and are used for NMF unmixing to figure out the underlying single-phase spectra, reducing the risk of identifying noisy features for NMF.

Figure 7 shows the unmixed 12 pseudo-spectra components 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. Note that cosine similarity is a metric to measure the similarity between two vectors (i.e., spectra in this case). Given two vectors $\mathbf{x}$ and $\mathbf{y}$, cosine similarity can be formulated as:

$$\text{cosine similarity}(\mathbf{x}, \mathbf{y}) = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

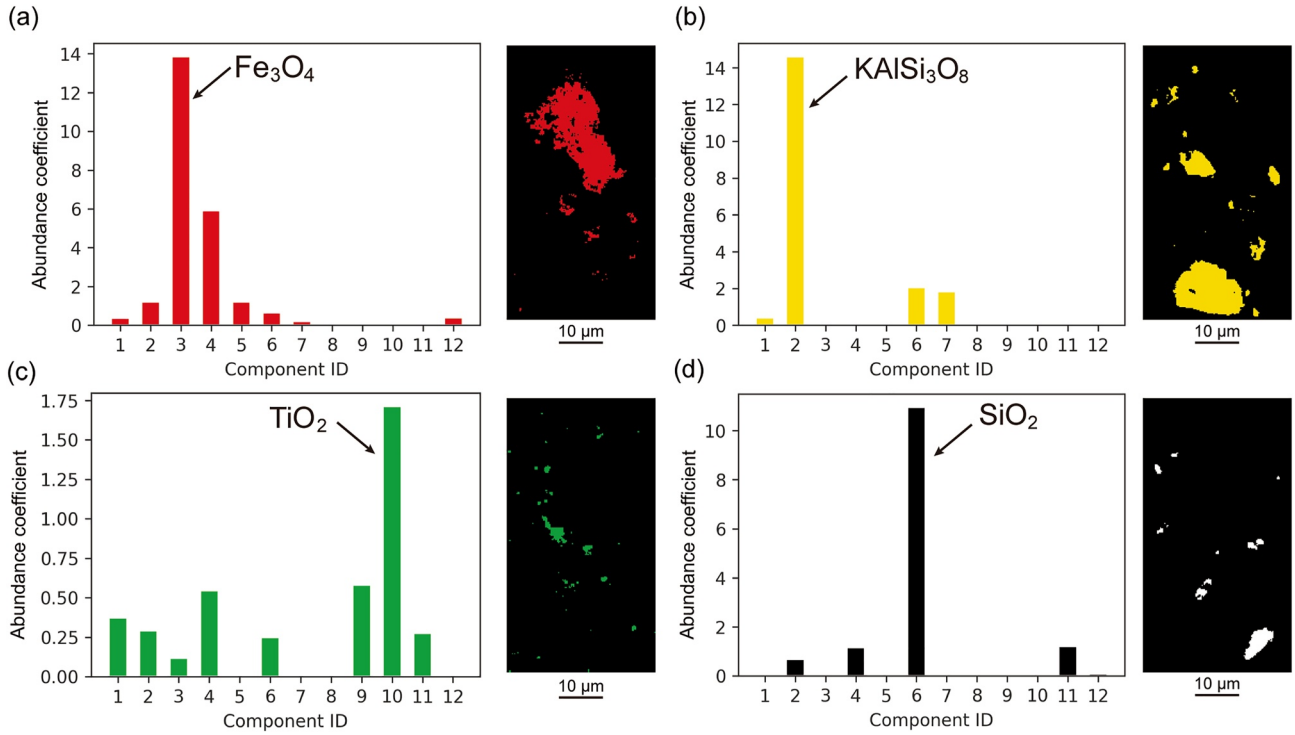

Figure 8. Bar charts of abundance coefficients and pixel distributions showing the importance of NMF components for each cluster. The underlying single-phase spectrum of each cluster of (a) Fe_3O_4, (b) $KAlSi_3O_8$, (c) TiO_2, and (d) SiO_2 can be identified according to the abundance coefficients.

where θ is the angle between the two vectors, $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ represent the L2 norm of $\mathbf{x}$ and $\mathbf{y}$. If the two given vectors are exactly the same, the cosine similarity will be 1 (or 100%); if they are decorrelated, the cosine similarity will be 0 (or 0%).

However, the unmixing process is not perfect. First, some phases are not properly unmixed, for example, TiO_2 and $ZnCO_3$ in components #10 and #12, respectively. This may result from the relatively small amount of these phases in the data set. 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, for example, 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 6). This leads to imbalanced signal intensity scales among different cluster-spectra, for example, 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, that is, 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 be interpreted as the general background, noise introduced by instrument artifacts, or noisy features due to the low-quality initial data set. 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 captures instrumental artifacts or noisy features. 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 8 shows the analysis of the abundance coefficients for each component, indicating the importance of the contribution of each component. As shown in Figure 8a, cluster #1 can be approximated using components #3 and #4 with an abundance coefficient of 13.9

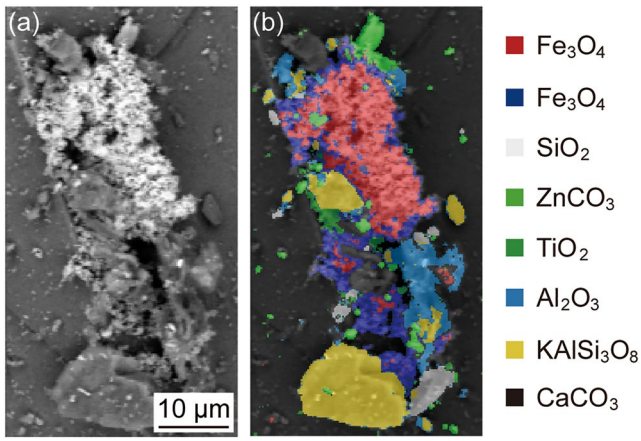

Figure 9. Backscattered electron image of the synthetic mixture data set and the corresponding phase map according to the NMF unmixing analysis. Note that red and dark blue represent the same phase of Fe_3O_4 but are unmixed from different clusters.

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, that is, only one or two components are dominant, as shown in Figures 8b–8d. 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 data set (Figure 9).

3.4. Benchmarking

We compare SIGMA with some other standard approaches with the synthetic mixture sample data set (Table 1). Two MSA methods are used as the baseline (i.e., independent component analysis (ICA, Hyvärinen & Oja, 2000; Seung & Lee, 2001)) and NMF, where the selection of the number of components for NMF/ICA is based on the Scree plot of PCA (Figure S1 in Supporting Information S1). Additionally, clustering directly on 9D pixels using GMM followed by unmixing using NMF (denoted as GMM-NMF) is also employed to demonstrate the importance of the dimensionality reduction step by autoencoder.

The two MSA methods have poorer performance in both phase identification and spectrum accuracy. ICA identifies 2 of 7 phases with an average 36.5% similarity (Figure S2 in Supporting Information S1), while NMF identifies 3 of 7 phases with an average 64.3% similarity (Figure S3 in Supporting Information S1). Two reasons are conceivable. First, the properties of non-negativity and statistical dependence of sources in HSI-EDS data sets may be the cause of failure of the ICA algorithms (Nascimento & Dias, 2005). Second, for ICA and NMF, low average counts per pixel of the data set (i.e., noisy EDS spectra) may pose a challenge for the algorithms to identify components of single-phase spectra. From the Scree plot (Figure S1 in Supporting Information S1), the substantial drop in the explained variance ratio of the first three principal components reflects the influence of the noisy input. Consequently, only three NMF/ICA components can be identified. Although a higher number of components can be assigned, it does not yield more meaningful components; they are similar, noisy spectra, instead. On the other hand, SIGMA identifies all phases (7 of 7) in the synthetic mixture sample and achieves better average similarity (83.0%) to the ground truth single-phase spectra than the common standard methods. However, without non-linear dimensionality reduction, GMM-NMF can only detect 6 of 7 phases and yield an average 72.9% similarity (Figure S4 in Supporting Information S1). The drop in performance demonstrates the importance of dimensionality reduction using autoencoder.

3.5. Testing SIGMA on Exhaust Pipe Residue Particulate Matter Data Set

We evaluate the performance of SIGMA on the exhaust pipe residue particulate matter data set, 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 informa-

Table 1
Benchmark Results Comparing SIGMA With Previous Standard Methods

Method	Number of phases identified	Average spectrum similarity to ground truth
ICA (Hyvärinen & Oja, 2000)	2 of 7	36.5%
NMF (Kotula et al., 2003)	3 of 7	64.3%
GMM-NMF	6 of 7	72.9%
SIGMA	**7 of 7**	**83.0%**

Note. The bold values hightlight the performance of SIGMA.

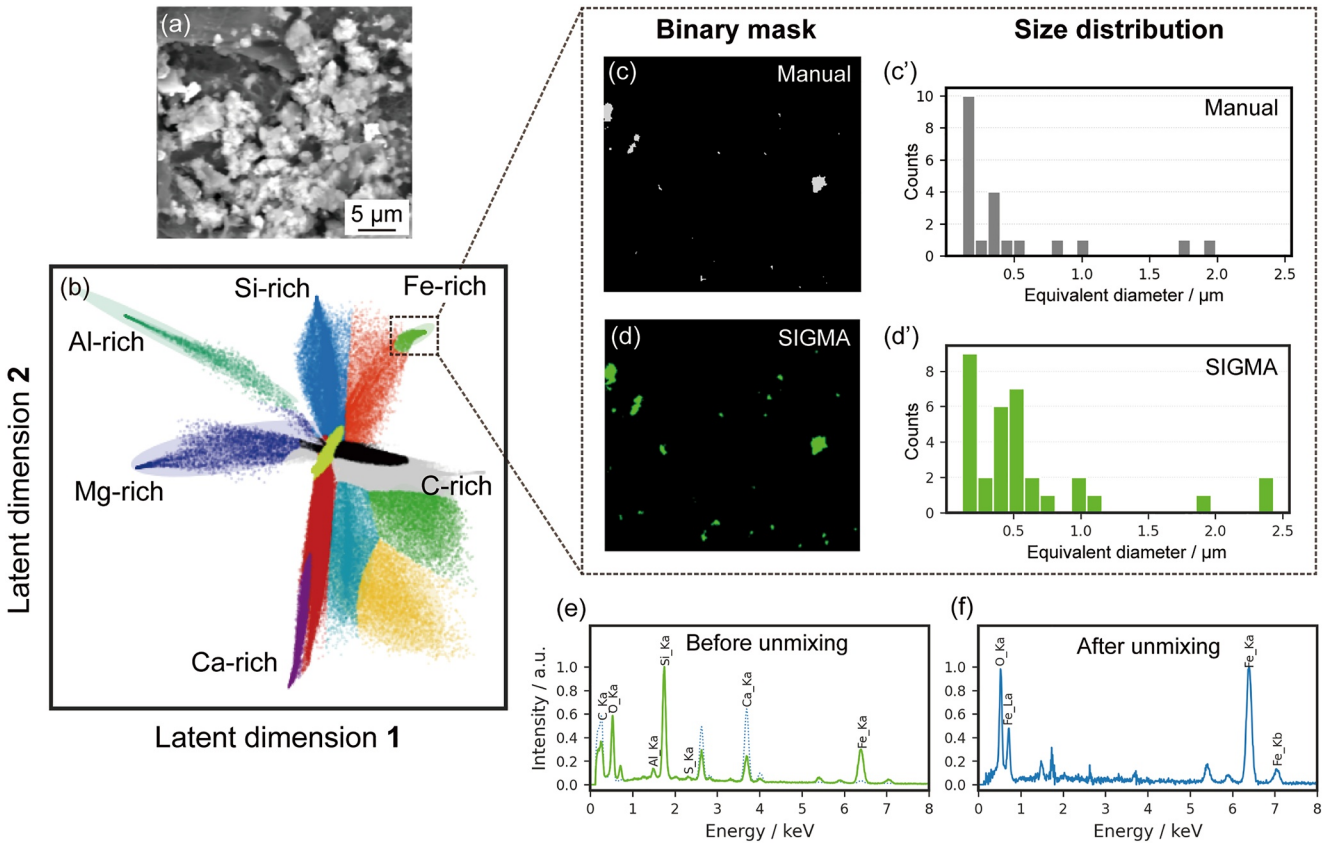

Figure 10. Phase identification and elemental signal unmixing on particulate matter data set using SIGMA. (a) Backscattered electron (BSE) image; (b) 2D latent space modeled by autoencoder showing the GMM clustering result, where clusters are marked with different colors and overlapped with the associated 95% confidence ellipses; (c) manually identified pixel distribution of Fe-bearing particles and (c′) associated size distribution; (d) pixel distribution in the Fe-rich cluster of the latent space and (d′) associated size distribution of the Fe-rich particles. Note that 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.

tion (Figure 10). Figure 10a shows the backscattered electron (BSE) image. After pre-processing and normalizing (the same procedure in the synthetic mixture data set), the data set is built into elemental intensity maps with the dimensions of 396 × 336-pixel × 8-spectral-channel (i.e., X-ray lines of O Kα, Fe Kα, Mg Kα, Ca Kα, Al Kα, C Kα, Si Kα, and S Kα). Then, an autoencoder is trained to learn the 2D representation of pixels. Figure 10b shows the autoencoder-modeled 2D latent space and the result of GMM clustering ($K = 13$), where data points 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). Figures 10c and 10c′ show the spatial distribution of Fe-rich pixels and the associated size distribution obtained by manual analysis (i.e., using a thresholding technique on the Fe intensity map); Figure 10d and 10d′ show the SIGMA analysis of the spatial distribution of Fe-rich pixels and the associated size distribution. SIGMA can not only identify most of the particles recognized by the manual analysis but also detect some overlooked particles that may be locations with multiple mineral particles stacked on each other. Prior to the unmixing step, although containing the Fe Kα peak well above the average, the sum spectrum (Figure 10e) 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 10f). 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.

4. 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) data set. Specifically, we apply non-linear dimensionality reduction to the HSI data set 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 data sets. For the known synthetic mixture data set, all seven major phases are identified and verified by the individually measured EDS spectra, revealing the accuracy (i.e., average cosine similarity = 83.0%) of our technique. For the particulate matter data set, 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 data sets, such as electron energy loss spectroscopy (EELS), scanning tunneling microscopy (STM), and time-of-flight secondary ion mass spectrometry (ToF-SIMS), providing a reliable analysis in a fully automated manner.

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.7114747).

Acknowledgments
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No 101005611. 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.

References

Adam, E., Mutanga, O., & Rugege, D. (2010). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review. *Wetlands Ecology and Management*, *18*(3), 281–296. https://doi.org/10.1007/s11273-009-9169-z

Afromowitz, M. A., Callis, J. B., Heimbach, D. M., DeSoto, L. A., & Norton, M. K. (1988). Multispectral imaging of burn wounds: A new clinical instrument for evaluating burn depth. *IEEE Transactions on Biomedical Engineering*, *35*(10), 842–850. https://doi.org/10.1109/10.7291

Aguiar, J., Gong, M. L., Unocic, R., Tasdizen, T., & Miller, B. (2019). Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Science Advances*, *5*(10), eaaw1949. https://doi.org/10.1126/sciadv.aaw1949

Antczak, K. (2018). Deep recurrent neural networks for ECG signal denoising. arXiv:10.48550/ARXIV.1807.11551.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv:10.48550/ARXIV.1607.06450.

Bellman, R., Corporation, R., & Collection, K. M. R. (1957). *Dynamic programming*. Princeton University Press.

Benhalouche, F. Z., Karoui, M. S., & Deville, Y. (2019). An NMF-based approach for hyperspectral unmixing using a new multiplicative-tuning linear mixing model to address spectral variability. In *Paper presented at 2019 27th European Signal Processing Conference (EUSIPCO), 2–6 September 2019*.

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., & Chanussot, J. (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. https://doi.org/10.1109/jstars.2012.2194696

Bishop, C. M., & Nasrabadi, N. M. (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. https://doi.org/10.1093/jxb/erl123

Blanco-Portals, J., Peiró, F., & Estradé, S. (2022). Strategies for EELS data analysis. Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. *Microscopy and Microanalysis*, *28*(1), 109–122. https://doi.org/10.1017/s1431927621013696

Bosman, M., Watanabe, M., Alexander, D. T. L., & Keast, V. J. (2006). Mapping chemical and bonding information using multivariate analysis of electron energy-loss spectrum images. *Ultramicroscopy*, *106*(11), 1024–1032. https://doi.org/10.1016/j.ultramic.2006.04.016

Carrasco, O., Gomez, R., Chainani, A., & Roper, W. (2003). *Hyperspectral imaging applied to medical diagnoses and food safety*. SPIE.

Chen, Z., Andrejevic, N., Drucker, N. C., Nguyen, T., Xian, R. P., Smidt, T., et al. (2021). Machine learning on neutron and X-ray scattering and spectroscopies. *Chemical Physics Reviews*, *2*(3), 031301. https://doi.org/10.1063/5.0049111

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Duan, X., Yang, F., Antono, E., Yang, W., Pianetta, P., Ermon, S., et al. (2016). Unsupervised data mining in nanoscale X-ray spectro-microscopic study of NdFeB magnet. *Scientific Reports*, *6*(1), 34406. https://doi.org/10.1038/srep34406

Durdziński, P. T., Dunant, C. F., Haha, M. B., & Scrivener, K. L. (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. https://doi.org/10.1016/j.cemconres.2015.02.008

Dusseldorp, A., Kruize, H., Brunekreef, B., Hofschreuder, P., De Meer, G., & Van Oudvorst, A. (1995). Associations of PM_{10} and airborne iron with respiratory health of adults living near a steel factory. *American Journal of Respiratory and Critical Care Medicine*, *152*(6), 1932–1939. https://doi.org/10.1164/ajrccm.152.6.8520758

Ede, J. M. (2020). Warwick electron microscopy datasets. *Machine Learning: Science and Technology*, *1*(4), 045003. https://doi.org/10.1088/2632-2153/ab9c3c

Ede, J. M. (2021). Deep learning in electron microscopy. *Machine Learning: Science and Technology*, *2*(1), 011004. https://doi.org/10.1088/2632-2153/abd614

Feng, Y.-Z., & Sun, D.-W. (2012). Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition*, *52*(11), 1039–1058. https://doi.org/10.1080/10408398.2011.651542

Funk, C. C., Theiler, J., Roberts, D. A., & Borel, C. C. (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. https://doi.org/10.1109/36.934073

Gendrin, C., Roggo, Y., & Collet, C. (2008). Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review. *Journal of Pharmaceutical and Biomedical Analysis*, *48*(3), 533–553. https://doi.org/10.1016/j.jpba.2008.08.014

Goldstein, J. I., Newbury, D. E., Michael, J. R., Ritchie, N. W., Scott, J. H. J., & Joy, D. C. (2017). *Scanning electron microscopy and X-ray microanalysis*. Springer.

Govender, M., Chetty, K., & Bulcock, H. (2007). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA*, *33*(2), 145–151. https://doi.org/10.4314/wsa.v33i2.49049

Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging – An emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, *18*(12), 590–598. https://doi.org/10.1016/j.tifs.2007.06.001

Han, Y., Jang, J., Cha, E., Lee, J., Chung, H., Jeong, M., et al. (2021). Deep learning STEM-EDX tomography of nanocrystals. *Nature Machine Intelligence*, *3*(3), 267–274. https://doi.org/10.1038/s42256-020-00289-5

He, X., Yan, S., Hu, Y., Niyogi, P., & Zhang, H.-J. (2005). Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 328–340. https://doi.org/10.1109/tpami.2005.55

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Scienc*e, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4), 411–430. https://doi.org/10.1016/s0893-6080(00)00026-5

Jany, B. R., Janas, A., & Krok, F. (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. https://doi.org/10.1021/acs.nanolett.7b01789

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Kannan, R., Ievlev, A. V., Laanait, N., Ziatdinov, M. A., Vasudevan, R. K., Jesse, S., & Kalinin, S. V. (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. https://doi.org/10.1186/s40679-018-0055-8

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:10.48550/ARXIV.1412.6980.

Kotula, P. G., Keenan, M. R., & Michael, J. R. (2003). Automated analysis of SEM X-ray spectral images: A powerful new microanalysis tool. *Microscopy and Microanalysis*, *9*(1), 1–17. https://doi.org/10.1017/s1431927603030058

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Li, X., Dyck, O. E., Oxley, M. P., Lupini, A. R., McInnes, L., Healy, J., et al. (2019). Manifold learning of four-dimensional scanning transmission electron microscopy. *npj Computational Materials*, *5*(1), 5. https://doi.org/10.1038/s41524-018-0139-y

Lin, Z., Chen, Y., Zhao, X., & Wang, G. (2013). Spectral-spatial classification of hyperspectral image using autoencoders. In *Paper presented at 2013 9th International Conference on Information, Communications & Signal Processing, 10–13 December 2013*.

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2020). Dying relu and initialization: Theory and numerical examples. *Communications in Computational Physics*, *28*(5), 1671–1706. https://doi.org/10.4208/cicp.oa-2020-0165

MacRae, C., Wilson, N., Torpy, A., Rossek, U., & Rohde, M. (2007). The holistic approach to data collection and hyperspectral analysis. *Microscopy and Microanalysis*, *13*(S2), 1358–1359. https://doi.org/10.1017/s1431927607071929

Maher, B. A., Ahmed, I. A. M., Karloukovski, V., MacLaren, D. A., Foulds, P. G., Allsop, D., et al. (2016). Magnetite pollution nanoparticles in the human brain. *Proceedings of the National Academy of Sciences*, *113*(39), 10797–10801. https://doi.org/10.1073/pnas.1605941113

Maher, B. A., González-Maciel, A., Reynoso-Robles, R., Torres-Jardón, R., & Calderón-Garcidueñas, L. (2020). Iron-rich air pollution nanoparticles: An unrecognised environmental risk factor for myocardial mitochondrial dysfunction and cardiac oxidative stress. *Environmental Research*, *188*, 109816. https://doi.org/10.1016/j.envres.2020.109816

Malinowski, E. R., & Howery, D. G. (1980). *Factor analysis in chemistry*. Wiley.

Martineau, B. H., Johnstone, D. N., van Helvoort, A. T. J., Midgley, P. A., & Eggeman, A. S. (2019). Unsupervised machine learning applied to scanning precession electron diffraction data. *Advanced Structural and Chemical Imaging*, *5*(1), 3. https://doi.org/10.1186/s40679-019-0063-3

McAuliffe, T. P., Dye, D., & Britton, T. B. (2020). Spherical-angular dark field imaging and sensitive microstructural phase clustering with unsupervised machine learning. *Ultramicroscopy*, *219*, 113132. https://doi.org/10.1016/j.ultramic.2020.113132

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

Molchanov, V. F., & Linsen, L. (2018). Overcoming the curse of dimensionality when clustering multivariate volume data. In *Paper presented at VISIGRAPP*.

Nascimento, J. M. P., & Dias, J. M. B. (2005). Does independent component analysis play a role in unmixing hyperspectral data? *IEEE Transactions on Geoscience and Remote Sensing*, *43*(1), 175–187. https://doi.org/10.1109/tgrs.2004.839806

Parish, C. M. (2019). Fuzzy clustering to merge EDS and EBSD datasets with crystallographic ambiguity. *Microscopy and Microanalysis*, *25*(S2), 134–135. https://doi.org/10.1017/s1431927619001405

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Potapov, P., & Lubk, A. (2019). Optimal principal component analysis of STEM XEDS spectrum images. *Advanced Structural and Chemical Imaging*, *5*(1), 4. https://doi.org/10.1186/s40679-019-0066-0

Roberts, G., Haile, S. Y., Sainju, R., Edwards, D. J., Hutchinson, B., & Zhu, Y. (2019). Deep learning for semantic segmentation of defects in advanced STEM images of steels. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-49105-0

Roels, J., & Saeys, Y. (2019). Cost-efficient segmentation of electron microscopy images using active learning. arXiv:10.48550/ARXIV.1911.05548.

Rossouw, D., Burdet, P., de la Peña, F., Ducati, C., Knappett, B. R., Wheatley, A. E. H., & Midgley, P. A. (2015). Multicomponent signal unmixing from nanoheterostructures: Overcoming the traditional challenges of nanoscale X-ray analysis via machine learning. *Nano Letters*, *15*(4), 2716–2720. https://doi.org/10.1021/acs.nanolett.5b00449

Rossouw, D., Knappett, B. R., Wheatley, A. E. H., & Midgley, P. A. (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. https://doi.org/10.1002/ppsc.201600096

Rui, X., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/tnn.2005.845141

Seung, D., & Lee, L. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, *13*, 556–562.

Sheikh, H. A., Maher, B. A., Karloukovski, V., Lampronti, G. I., & Harrison, R. J. (2022). Biomagnetic characterization of air pollution particulates in Lahore, Pakistan. *Geochemistry, Geophysics, Geosystems*, *23*(2), e2021GC010293. https://doi.org/10.1029/2021gc010293

Shlens, J. (2014). A tutorial on principal component analysis. arXiv:10.48550/ARXIV.1404.1100.

Stork, C. L., & Keenan, M. R. (2010). Advantages of clustering in the phase classification of hyperspectral materials images. *Microscopy and Microanalysis*, *16*(6), 810–820. https://doi.org/10.1017/s143192761009402x

Sumithra, V., & Surendran, S. (2015). A review of various linear and non linear dimensionality reduction techniques. *International Journal of Computer Science and Information Technology*, *6*, 2354–2360.

Tao, C., Pan, H., Li, Y., & Zou, Z. (2015). Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(12), 2438–2442. https://doi.org/10.1109/lgrs.2015.2482520

Tao, X., Paoletti, M. E., Han, L., Wu, Z., Ren, P., Plaza, J., et al. (2022). A new deep convolutional network for effective hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 6999–7012. https://doi.org/10.1109/jstars.2022.3200733

Teng, C., & Gauvin, R. (2020). Multivariate statistical analysis on a SEM/EDS phase map of rare Earth minerals. *Scanning*, *2020*, 2134516. https://doi.org/10.1155/2020/2134516

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, *61*(3), 611–622. https://doi.org/10.1111/1467-9868.00196

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *Paper presented at 2015 IEEE Information Theory Workshop (ITW)*. IEEE.

Urakubo, H., Bullmann, T., Kubota, Y., Oba, S., & Ishii, S. (2019). UNI-EM: An environment for deep neural network-based automated segmentation of neuronal electron microscopic images. *Scientific Reports*, *9*(1), 19413. https://doi.org/10.1038/s41598-019-55431-0

Vasudevan, R. K., Laanait, N., Ferragut, E. M., Wang, K., Geohegan, D. B., Xiao, K., et al. (2018). Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images. *npj Computational Materials*, *4*(1), 30. https://doi.org/10.1038/s41524-018-0086-7

Vekemans, B., Vincze, L., Brenker, F. E., & Adams, F. (2004). Processing of three-dimensional microscopic X-ray fluorescence data. *Journal of Analytical Atomic Spectrometry*, *19*(10), 1302–1308. https://doi.org/10.1039/b404300f

Wang, H., Wang, W., Yang, J., & Yu, P. S. (2002). Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data* (pp. 394–405). Association for Computing Machinery.

Wit, E., Heuvel, E. V. D., & Romeijn, J.-W. (2012). 'All models are wrong…': An introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217–236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

Yan, B., McJunkin, T. R., Stoner, D. L., & Scott, J. R. (2006). Validation of fuzzy logic method for automated mass spectral classification for mineral imaging. *Applied Surface Science*, *253*(4), 2011–2017. https://doi.org/10.1016/j.apsusc.2006.03.093

Yann, L., & Ishan, M. (2021). Self-supervised learning: The dark matter of intelligence. In *Meta AI Blog*.

Yokoyama, Y., Terada, T., Shimizu, K., Nishikawa, K., Kozai, D., Shimada, A., et al. (2020). Development of a deep learning-based method to identify "good" regions of a cryo-electron microscopy grid. *Biophysical Reviews*, *12*(2), 349–354. https://doi.org/10.1007/s12551-020-00669-6

Yoon, D., Lim, H. S., Jung, K., Kim, T. Y., & Lee, S. (2019). Deep learning-based electrocardiogram signal noise detection and screening model. *Healthcare informatics research*, *25*(3), 201–211. https://doi.org/10.4258/hir.2019.25.3.201

Yu, Z. X., Wei, S. C., Zhang, J. W., Wang, B., Wang, Y. J., Liang, Y., & Tian, H. L. (2020). High-throughput, algorithmic determination of pore parameters from electron microscopy. *Computational Materials Science*, *171*, 109216. https://doi.org/10.1016/j.commatsci.2019.109216

Zaefferer, S., & Habler, G. (2017). Scanning electron microscopy and electron backscatter diffraction. In W. Heinrich & R. Abart (Eds.), *Mineral reaction kinetics: Microstructures, textures, chemical and isotopic signatures*. European Mineralogical Union and Mineralogical Society of Great Britain and Ireland.