Deep Learning-Assisted Multivariate Analysis for Nanoscale Characterization of Heterogeneous Beam-Sensitive Materials

Felix Utama Kosasih, Fanzhi Su, Tian Du, Sinclair Ryley Ratnasingham, Joe Briscoe, Caterina Ducati

Microscopy and Microanalysis, 2023, **00**, 1–15
https://doi.org/10.1093/micmic/ozad033

Microscopy AND Microanalysis

Original Article

Deep Learning-Assisted Multivariate Analysis for Nanoscale Characterization of Heterogeneous Beam-Sensitive Materials

Felix Utama Kosasih[1,†], Fanzhi Su[1], Tian Du[2,‡], Sinclair Ryley Ratnasingham[2,3], Joe Briscoe[2], and Caterina Ducati[1,*]

[1]*Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, UK*
[2]*School of Engineering and Materials Science and Materials Research Institute, Queen Mary University of London, London E1 4NS, UK*
[3]*Department of Materials and Centre for Processable Electronics, Molecular Science Research Hub, Imperial College London, London W12 0BZ, UK*

*Corresponding author: Caterina Ducati, E-mail: cd251@cam.ac.uk

[†]Current address: Energy Research Institute @ NTU, Nanyang Technological University, 50 Nanyang Drive, Singapore 637553.

[‡]Current address: Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander University Erlangen-Nürnberg, Martensstraße 7, 91058 Erlangen, Germany.

Abstract

Nanoscale materials characterization often uses highly energetic probes which can rapidly damage beam-sensitive materials, such as hybrid organic–inorganic compounds. Reducing the probe dose minimizes the damage, but often at the cost of lower signal-to-noise ratio (SNR) in the acquired data. This work reports the optimization and validation of principal component analysis (PCA) and nonnegative matrix factorization for the postprocessing of low-dose nanoscale characterization data. PCA is found to be the best approach for data denoising. However, the popular scree plot-based method for separation of principal and noise components results in inaccurate or excessively noisy models of the heterogeneous original data, even after Poissonian noise weighting. Manual separation of principal and noise components produces a denoised model which more accurately reproduces physical features present in the raw data while improving SNR by an order of magnitude. However, manual selection is time-consuming and potentially subjective. To suppress these disadvantages, a deep learning-based component classification method is proposed. The neural network model can examine PCA components and automatically classify them with an accuracy of >99% and a rate of ~2 component/s. Together, multivariate analysis and deep learning enable a deeper analysis of nanoscale materials' characterization, allowing as much information as possible to be extracted.

Key words: convolutional neural network, energy-dispersive X-ray spectroscopy, materials' characterization, principal component analysis, scanning transmission electron microscopy

Introduction

Materials' characterization is an essential activity to understand the relationship between a material's structure, properties, processing, and performance (Burnett & Withers, 2019). As nanomaterials and nanostructured devices become increasingly relied upon for various applications in our daily life, the importance of nanoscale materials' characterization has risen accordingly. Many characterization techniques can reach a spatial resolution in the nanometer scale by using a tightly focused probe beam composed of photons, ions, neutrons, protons, or electrons. To generate a high signal count from the specimen, such a beam must contain a sufficient number of probe particles within its narrow diameter. This parameter is called dose, defined as the number of probe particles per unit of beam area and expressed in units of particle/nm^2. If the dose is too low, the signal-to-noise ratio (SNR) in the obtained data might also be too low and the measurement uncertainty too high, rendering the data statistically insignificant.

This minimum dose requirement becomes an impediment when the material under scrutiny is easily damaged by the probe particles (beam-sensitive), such as metal–organic frameworks or organic–inorganic hybrid halide perovskites (Chen et al., 2020; Ilett et al., 2020). In such materials, the energy deposited by an intense probe beam through inelastic probe particle-matter scattering is likely enough to induce decomposition and structural degradation (Henderson, 1995). For this reason, when working on beam-sensitive materials, one cannot simply increase the probe dose to obtain a higher signal count. Rather, the possible solutions are more efficient signal detectors or postprocessing of the data to improve its SNR (Schlossmacher et al., 2010; Mohan et al., 2020; Zhang et al., 2020). Of these two, the latter is almost always cheaper and more accessible. However, postprocessing may also introduce artifacts or obscure genuine features in the data if it is not properly executed or otherwise unsuitable for the data type. Therefore, postprocessing algorithms must be carefully applied to the raw data to ensure that any extracted conclusions are valid.

A particularly powerful approach for nanoscale characterization of beam-sensitive materials is to combine spectrum imaging and postprocessing by multivariate analysis (MVA). Spectrum imaging is the acquisition of three-dimensional (3D) data cubes where two of the axes (navigation axes) correspond to a two-dimensional (2D) region of interest on the

Received: April 12, 2022. **Revised:** November 17, 2022. **Accepted:** March 3, 2023

specimen, while the other (signal axis) represents a list of acquired data. As the probe scans through the 2D area, an energy spectrum, diffractogram, or other data types is obtained from each probed spot, stored in the signal axis, and associated with the corresponding pixel in the navigation axes. In nanoscale characterization with modern instruments, the number of data points in a single spectrum image can easily reach the order of 10^8. Although each of these data points is acquired separately, they are very likely to share similar characteristics with one another. For example, parts of the specimen that are mapped in adjacent pixels have a high likelihood of having the same composition and structure (surfaces and interfaces being the obvious exceptions). The high number of related data points makes spectrum images ideal for processing with MVA algorithms such as principal component analysis (PCA) and nonnegative matrix factorization (NMF), which work by finding relationships between variables and leveraging them to extract meaningful patterns and information from a dataset.

The general principles of PCA and NMF are well known and are summarized in the Supporting Information. Briefly, PCA finds a set of orthogonal components (linear combinations of the data's variables) which account for most of the original data's variance (Pearson, 1901). These are called principal components, while all the other components are considered noise and discarded. NMF works similarly but replaces the constraint of orthogonality with nonnegativity (Paatero & Tapper, 1994; Lee & Seung, 1999). Despite their wide use, there are subtle nuances involved in their application which deserve further refinement. For example, the discrimination between principal and noise components in PCA is normally performed automatically using the elbow point in a scree plot (Supplementary Fig. 1a). While this method is straightforward in principle, it may not necessarily be easy or accurate in practice. Then, if the elbow point approach is indeed nonideal, another method may be needed to identify the principal components more accurately. Furthermore, the suitability of the elbow point to choose the number of output NMF components should also be assessed. Finally, for the purpose of denoising, it is not always clear whether the denoised dataset should be constructed using PCA or NMF components to model the original data most closely and obtain accurate analysis results.

In this article, we report the optimization and validation of PCA and NMF for the postprocessing of multidimensional materials' characterization data acquired with minimum probe dose. Specifically, PCA and NMF were performed on nanoscale cross-sectional energy-dispersive X-ray spectroscopy (EDX) data of organic–inorganic hybrid halide perovskite solar cells (PSCs) acquired in a scanning transmission electron microscope (STEM). This dataset was chosen due to the massive research interest on PSCs and because it is well known that halide perovskites contain nanoscale compositional heterogeneities which affect their optoelectronic properties and stability (Tennyson et al., 2019; Chen & Zhou, 2020; Doherty et al., 2020). The recent advent of halide perovskites containing up to seven different ions, combined with the ease of ion migration in halide perovskites, provide additional motivation to understand the compositional landscape of PSCs (Azpiroz et al., 2015; Eames et al., 2015; Haruyama et al., 2015; Saliba et al., 2016). The stages of MVA optimization performed in this work include (a) proper discrimination between principal and noise components in PCA, (b) choosing the right number of output components for NMF, and (c) comparison between denoised models constructed from PCA or NMF components. Of the two investigated algorithms, we find that PCA is the best option for data denoising, while NMF works well to decompose the data into components easily attributable to physical features. Furthermore, we show that the elbow point method produced PCA-denoised models that are either inaccurate or unnecessarily noisy when the specimens are highly heterogeneous. Rather, manual separation between the principal and noise components produce denoised models which faithfully replicate features contained in the original data while maximizing SNR. However, manual sorting through hundreds of PCA components can be very laborious, time-intensive, and can potentially introduce subjectivity into the component selection. We solve this problem by demonstrating how deep learning-based convolutional neural network (CNN) can be used to automate the selection of principal components. Our CNN workflow can accurately identify principal components in a fast and bias-free manner, resulting in optimum data denoising.

The powerful combination of MVA algorithms and CNN can help us sieve through the vast amounts of data produced in nanoscale characterization through denoising or decomposition into physically meaningful components. In the context of beam-sensitive materials, having an automated and reliable data denoising is especially invaluable as it enables data acquisition using far lower probe doses than what would be required otherwise, and thus reduces specimen damage (Kosasih et al., 2020).

Materials and Methods

Perovskite Solar Cell Fabrication

The precursor solution for $CH_3NH_3PbI_3$ devices was prepared by dissolving equimolar concentrations (1.5 mol/dm^3) of PbI_2 and CH_3NH_3I in a mixed solvent of *N*, *N*-dimethylformamide (DMF) and dimethylsulfoxide (9:1.1, volume ratio). The solution was stirred at 60°C for 1 h and was passed through a 0.45 μm polytetrafluoroethylene filter before use; 40 μL precursor solution was dropped onto each substrate and spun at 4,000 RPM for 30 s. At the 7th second, 0.5 mL diethyl ether was dripped onto the spinning substrate. The control $CH_3NH_3PbI_3$ films were then annealed on a hot plate at 100°C for 20 min. The aerosol-treated $CH_3NH_3PbI_3$ films were preannealed at 100°C for 2 min to dry most of the solvent prior to aerosol treatment. Films were then placed within the preheated reactor, with the temperature set at 100°C. The treatment was carried out by flowing aerosolized DMF into the reactor at 0.5 dm^3/min for 5 min. The aerosol was generated using a piezoelectric generator. The substrates were placed in the central section of the reactor, approximately 4 cm from the aerosol inlet. After 5 min have elapsed, the aerosol flow was switched to N_2 and the samples were left on the heated graphite block for a further 5 min at the same temperature to sweep out the remaining DMF in the chamber. Samples were left to cool, then placed into a glovebox for thermal annealing at 100°C for 20 min.

All devices were fabricated on ITO-coated glass substrates sequentially cleaned in acetone, isopropanol, and deionized water (using ultrasonics) for 10 min followed by an N_2 dry. Prior to deposition, the substrates were treated by oxygen plasma for 10 min. Poly(N,N'-bis-4-butylphenyl-N,N'-bisphenyl)benzidine (PolyTPD, 0.25 wt% in chlorobenzene) was spin coated onto the ITO at 5,000 RPM for 20 s as the hole transport layer. After drying for 1 min, Poly[(9,9-bis(3′-((N,N-dimethyl)-N-ethylammonium)-propyl)-2,7-fluorene)-alt-2,7-(9,9-dioctylfluorene)] (PFN-Br, 0.05 wt% in methanol) was spin coated onto the hole transport layer at 5,000 RPM for 15 s as an interfacial modifier to reduce surface hydrophobicity. Solutions of the electron transport layer were prepared by dissolving 30 mg/mL phenyl-C_{61}-butyric acid methyl ester (PCBM) in chlorobenzene. The solution was stirred at 40°C for 1 h and filtered through a 0.45 μm polytetrafluoroethylene filter before use. The PCBM solution was spin coated on to $CH_3NH_3PbI_3$ films at 2,000 RPM for 45 s. An ultra-thin interfacial dipole layer was prepared by spin coating a bathocuproine solution (0.5 mg/mL in methanol) on top of the PCBM layer at 4,000 RPM for 30 s. Finally, the devices were completed by thermally evaporating 100 nm of Cu at a rate of 1 Å/s and a base pressure of 5×10^{-6} mbar.

Scanning Transmission Electron Microscopy

Cross-sectional lamellae of the PSCs were prepared with an FEI Helios Nanolab Dualbeam FIB/SEM following a standard procedure described elsewhere (Kosasih et al., 2020, 2022). The lamellae were immediately transferred into an FEI Tecnai Osiris STEM, minimizing air exposure to ~2 min. The STEM was operated at a 200 kV acceleration voltage and fitted with a Bruker Super-X silicon drift detector system for acquisition of EDX spectroscopy data, with a total collection solid angle of ~0.9 sr. STEM images were acquired in high-angle annular dark field mode using a Fischione detector, with a beam current of ~250 pA and a dwell time of 1 μs/pixel. STEM–EDX data was obtained with a beam current of ~140 pA, a dwell time of 40 ms/pixel, a spectral resolution of 5 eV/channel, and a spatial sampling of 10 nm/pixel as previously optimized elsewhere (Kosasih et al., 2020). PCA and NMF of STEM–EDX data was performed in HyperSpy, an open source Python package for multidimensional data analysis (de la Peña et al., 2020). STEM–EDX data was spectrally rebinned to a resolution of 20 eV/channel prior to application of PCA and NMF to improve the SNR. Prior to PCA, Poissonian noise in the data was normalized using a weighting algorithm developed by Keenan & Kotula (2004). PCA was executed using the singular value decomposition algorithm on the weighted data (HyperSpy, n.d.*a*, n.d.*b*). In HyperSpy, NMF was performed using scikit-learn's default NMF algorithm (Scikit Learn, n.d.). Quantitative elemental analysis of the denoised models was also performed in HyperSpy using the Cliff–Lorimer method (Cliff & Lorimer, 1975). The MVA code may be found at https://github.com/FanzhiSu/Deep-Learning-assisted-Multivariate-Analysis.git.

Image PreProcessing for Deep Learning Model

The preprocessing procedure started with manual classification of the first 100 PCA component scores into principal and noise ones. These 100 PCA scores of size 652 × 988 pixels were then de-stacked to 400 score images of smaller sizes, corresponding to the four EDX scan areas per stack as shown in Figure 1. This was done to address the data insufficiency problem by increasing the dataset size and also to show that the model's utility is not limited only to data stacks composed of dissimilar specimens. Score images in the minority class (principal components) were over-sampled by 14 times through offline data augmentation by randomly changing the image brightness and grayscale, horizontal and vertical flipping, and random rotations by 45°. Subsequently, the balanced dataset was resized to size 100 × 100 pixels to accelerate the training process and equalize the image dimensions. The resized images were normalized over the original dataset to stabilize the training process and further augmented on-the-fly with random cropping of the images to size 80 × 80 pixels. The preprocessed dataset was divided into training and testing sets with a ratio of 7:3.

Neural Network Architecture

The neural network consists of five hidden layers with a fully convolutional network as the backbone (see also Supplementary Fig. 8). It includes the following layers in order: 3 × 3 2D convolutional layer accepting 3 color channels and outputting 10 color channels, Max-Pool layer, rectified linear unit (ReLU) activation function, 3 × 3 2D convolutional layer accepting 10 color channels and outputting 200 color channels, Max-Pool layer, ReLU activation function, and 2 fully connected dense layers with softmax activation function for classification and extracting probability. The input dimension was set as 80 × 80. The network was trained with Adam optimizer and a learning rate of 5×10^{-4} for 10 iterations with a batch size of 10. Cross entropy loss was used as the loss function. All network training was conducted on a single Nvidia GeForce RTX 3070 GPU. The relevant code may be found at https://github.com/FanzhiSu/Deep-Learning-assisted-Multivariate-Analysis.git.

Neural Network Evaluation

To evaluate the CNN model's prediction accuracy, multiple evaluation metrics were employed:

(a) True positive (TP): the model correctly predicts a positive class (1);
True negative (TN): the model correctly predicts a negative class (0);
False positive (FP): the model predicts a positive class (1) by mistake;
False negative (FN): the model predicts a negative class (0) by mistake.

(b) Accuracy: the fraction of correct predictions over total predictions. It ranges from 0 to 1, with higher values indicating a better model. The accuracy can be calculated as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

(c) Sensitivity and specificity: these are two parameters used to evaluate model overfitting. They range from 0 to 1, with higher values indicating a better model.
Sensitivity: the probability that the model predicts a positive class (1) given that the case is actually a positive class:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

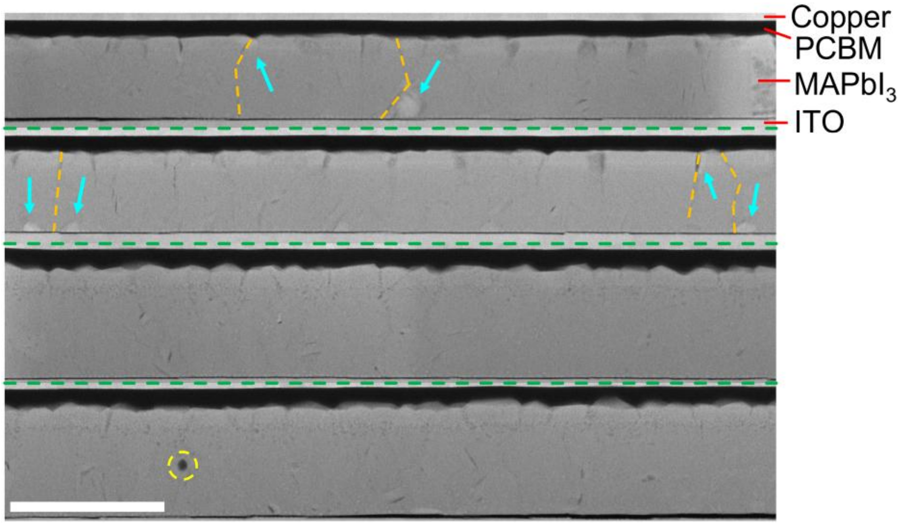


Fig. 1. A stack of four high angle dark field STEM images showing where the STEM–EDX scans were performed for each PSC cross section. The top (bottom) two scan areas are from the control (aerosol-treated) PSC. This arrangement of scan areas is used throughout this article. Dashed green lines mark the boundaries between scan areas. Dashed orange lines highlight some of the perovskite grain boundaries. Cyan arrows mark grains of a secondary phase in the perovskite film. Dashed yellow circle marks a damaged area where the focused electron beam was unintentionally parked for a few seconds. Scale bar represents 1 μm.

Specificity: the probability that the model predicts a negative class (0) given that the case is actually a negative class:

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

(d) Positive predicted value (PPV): the probabiliy that the model predicts a positive class (1) correctly:

$$\text{PPV} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

(e) Negative predicted value (NPV): the probabiliy that the model predicts a negative class (0) correctly:

$$\text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}}$$

PPV and NPV range from 0 to 1, with higher values indicating a better model.

Results and Discussion

In this study, two PSCs are used as a case study to find out how PCA and NMF can be best applied for data analysis. The first device is a spin-coated $CH_3NH_3PbI_3$ ($MAPbI_3$) PSC (control) and the second is a spin-coated $MAPbI_3$ PSC which was subsequently exposed to aerosolized DMF (aerosol-treated) as described in the Materials and Methods section. The scan areas mapped in this stack are shown in Figure 1, with the boundaries between scan areas marked by dashed green lines. The top two areas are from the control specimen, while the bottom two areas come from the aerosol-treated specimen. Each scan area contains four major layers, namely tin-doped indium oxide (ITO), $MAPbI_3$, PCBM, and copper.

The spin coating step was performed in exactly the same manner for both devices. Importantly, an annealing step was performed after spin coating for the control device, but it was delayed until after the aerosol treatment for the aerosol-treated device (Du et al., 2021). Previous research has shown that an annealing step of the same duration and temperature leads to CH_3NH_3I (MAI) loss through the perovskite grain boundaries, resulting in PbI_2 formation in the vicinity of those grain boundaries (Du et al., 2017). This also happened in the control device examined here, as indicated by the presence of bright grains next to the perovskite grain boundaries (cyan arrows and dashed orange lines in the top half of Fig. 1). These bright grains are shown below to be PbI_2. On the other hand, those grains are absent in the aerosol-treated device (bottom half of Fig. 1). This is because the aerosol treatment fostered further perovskite grain growth through solvent vapor-assisted Ostwald ripening, leading to a lower concentration of grain boundaries (Du et al., 2021). MAI loss and PbI_2 formation were thus suppressed. Consequently, the aerosol-treated device exhibits a thicker perovskite film and should be richer in C and N relative to the control sample due to the inhibited MAI loss. If properly constructed (using either PCA or NMF), a denoised STEM–EDX model should show the presence of PbI_2 in the control device and the difference in C and N concentrations between both devices. Therefore, these two features are used to evaluate the suitability of PCA and NMF for multidimensional data postprocessing.

We firstly used focused ion beam milling to extract four electron-transparent cross-sectional lamellae from the two PSCs. Then, we acquired STEM–EDX spectrum images from these lamellae and processed them in HyperSpy, an open source Python library for multidimensional data analysis (de la Peña et al., 2020). Before PCA can be performed on the spectrum images, it is necessary to apply a noise scaling operation on them. This is because EDX data acquisition is essentially a particle counting procedure, meaning the distribution of its measurement noise is Poissonian. Meanwhile, many PCA algorithms assume a Gaussian noise distribution in the data. Therefore, we firstly processed the spectrum images using a weighting procedure developed by Keenan & Kotula (2004) to ensure that our data is amenable to PCA. This Poissonian noise weighting method is appropriate to use as neither the mean spectrum nor the mean image of our dataset are sparse (Supplementary Fig. 2). Then, PCA and NMF are

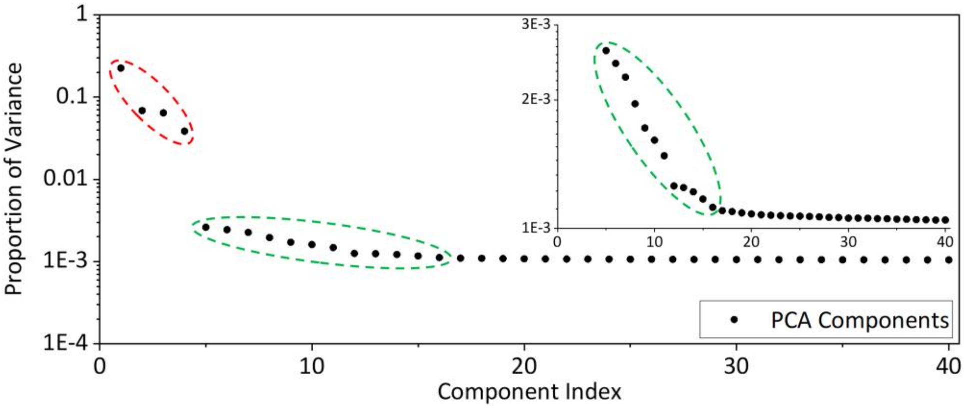


Fig. 2. Scree plot from the PCA procedure applied to the stack of STEM–EDX spectrum images. The inset shows a zoomed-in version where the second elbow point becomes clearly apparent. Dashed red ellipse marks the first four components appearing before the first elbow point. Dashed green ellipses mark components 5–16 appearing after the first but before the second elbow point.

performed in HyperSpy as well, using the singular value decomposition algorithm for PCA.

Determination of Principal Components from PCA and NMF

The most critical step in PCA denoising is the selection of components used to build the denoised model. The conventional approach is to use a scree plot, with components appearing before the elbow point considered as principal components and included in the model, while the rest are considered noise and discarded (Fig. 2; Supplementary Fig. 1a; Cattell, 1966; Zhu & Ghodsi, 2006). The primary advantage of this approach is the removal of bias or subjectivity in component choice, as a scree plot ranks components entirely based on statistical variance. However, this method is also more of an empirical guidance than an exact rule. For instance, it is reasonable to suspect that the elbow point method will eventually fail as specimens get more heterogeneous. This is because genuine spectral features which are only present in a few pixels may contribute a smaller portion of the data's overall variance compared to spectral noise which appear in many pixels, or which is related to very intense X-ray peaks. To provide an example in the context of PSCs, the component which represents small, localized inclusions such as PbI_2 grains is likely to be ranked lower in the scree plot than components showing noise from the organic transport layer's C–K_α peak or the electrode's metal peaks. Therefore, it is of great interest to investigate whether the elbow point approach can be used to accurately identify components containing real features for heterogeneous specimens.

PCA was performed on the stack of STEM–EDX spectrum images described above, and the resulting components were ranked in the scree plot shown in Figure 2. At first glance, selecting the principal components seems straightforward, as the first four components (marked by a dashed red ellipse in Fig. 2) clearly have far higher variances than the rest. Furthermore, it looks like the variance quickly plateaus after component 4. Therefore, the standard elbow point method identifies four principal components and discards the rest as noise. However, zooming in on the scree plot (inset in Fig. 2) reveals a secondary elbow point where components 5–16, marked by the dashed green ellipses in Figure 2, also show relatively high variances.

The distribution maps and spectra of components 1–4 are displayed in Figure 3. As expected from their very high variance, they show genuine physical features from the specimen and are indeed principal components. On the other hand, components 5–16 are a mix of real features and noise. Components 8 and 10 (Fig. 4) appear to correlate with a compositional difference between the control and aerosol-treated specimens, which is exactly the kind of useful information one would wish to find out by performing nanoscale characterization. Component 16 (Fig. 4) contains information on nanoscale heterogeneity as it shows small areas in the perovskite layer which are Pb-rich and I-poor compared to their surroundings. Components 8, 10, and 16 should be classified as principal components since their scores clearly have a nonrandom distribution of values and their loadings feature peaks and valleys whose energies correspond to X-ray lines of relevant elements. In contrast, the scores of components 5–7, 9, and 11–15 only show random distributions and their loadings are dominated by spikes instead of peaks and valleys (Supplementary Fig. 3). Therefore, these components can rightfully be considered noise. Importantly, principal components 8, 10, and 16 are interspersed with noise components 5–7, 9, and 11–15 in the scree plot. This shows that principal component selection should not be based entirely on variance ranking.

Normally, the number of principal components identified by the standard elbow point method in PCA (4 in this case) is fed into the NMF algorithm as p, or the desired number of output NMF components. However, the preceding discussion has shown that (a) there may be a secondary elbow point in the scree plot and (b) principal and noise components may be mixed between the first and second elbow points. Therefore, the next step is to examine the product of the NMF procedure while varying p from 4 to 16. Figure 5 shows the resulting components when $p = 4$. All four components show real features of the specimens, with scores that correlate well with the dark field images (Fig. 1) and loadings showing peaks attributable to X-ray lines. Components 1–4 show the perovskite layer (including the heterogeneities within), the copper electrode, the ITO layer, and the PCBM layer, respectively. Components 1 and 3 also show copper peaks from the copper sample half-grid. It is clear that NMF loadings are more directly

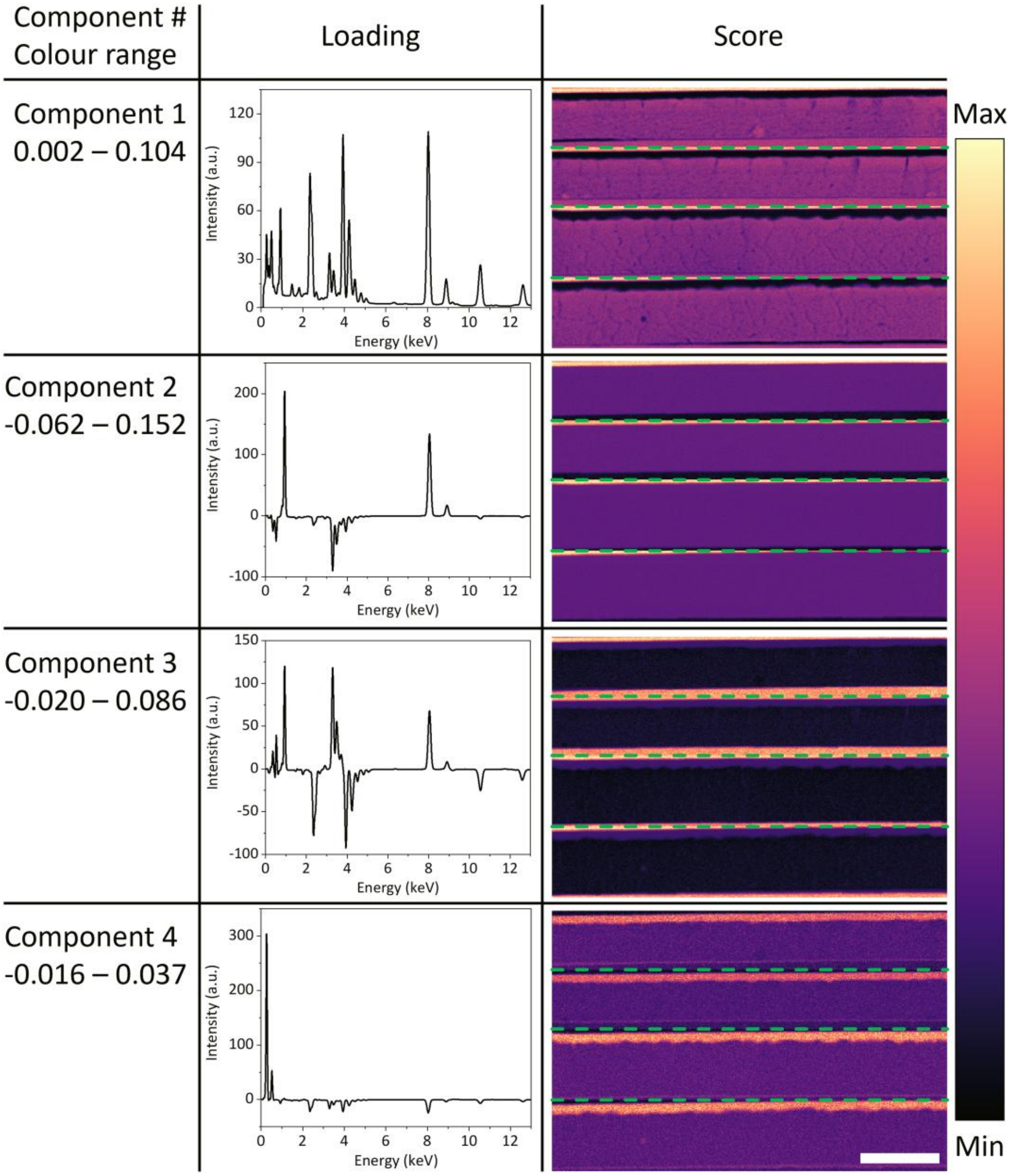


Fig. 3. PCA principal components (1–4) appearing before the first elbow point in the scree plot shown in Figure 2. These components represent genuine physical features in the specimen. Dashed green lines mark the boundaries between scan areas. Scale bar represents 1 μm.

comparable to the raw EDX data than PCA components. Both peak positions and peak intensities can be readily interpreted and assigned to specific X-ray lines.

When p is set to 16, only the first four components show physical features, while the rest consists of mostly noise. These four components (Fig. 6) appear similar to those shown in Figure 5, but there are some important differences. Most notably, the peaks attributable to Cu–K$_\alpha$ (8.04 keV) and K$_\beta$ (8.90 keV) in components 1–3 are greatly diminished, the O–K$_\alpha$ peaks (0.53 keV) disappear from the ITO and PCBM components, and the C–K$_\alpha$ signal (0.27 keV) in component 4 changed from a typical EDX peak into a very narrow spike. These changes occur because the signal corresponding to those peaks were assigned to the other 12 components instead, which can be grouped into three types as shown in Supplementary Figure 4. The first type (component 10, Supplementary Fig. 4) includes the Cu–K$_\alpha$ and K$_\beta$ signal missing from components 1–3. The second (components 5–7, 9, 11, 12, Supplementary Fig. 4) are sharp C–K$_\alpha$ spikes at energies surrounding 0.27 keV, suggesting that the C–K$_\alpha$ peak's signal count was split into several components. Finally, the third type (components 8, 13–15, Supplementary Fig. 4) contains the O–K$_\alpha$ signal which should have been present in the ITO and PCBM components. Comparing the components produced by $p = 4$ and $p = 16$, the former produces the more physically meaningful set of results because of the differences described above. For example, there is no physical basis for the separation of the O–K$_\alpha$ peak from the ITO and PCBM components. Indeed, its absence from those components when $p = 16$ (Fig. 6) means the loadings of those components no longer accurately represent ITO and PCBM, both of which are known to contain oxygen. Therefore, it can be concluded that the standard elbow point method is still useful to infer the most appropriate p for NMF from a PCA scree plot.

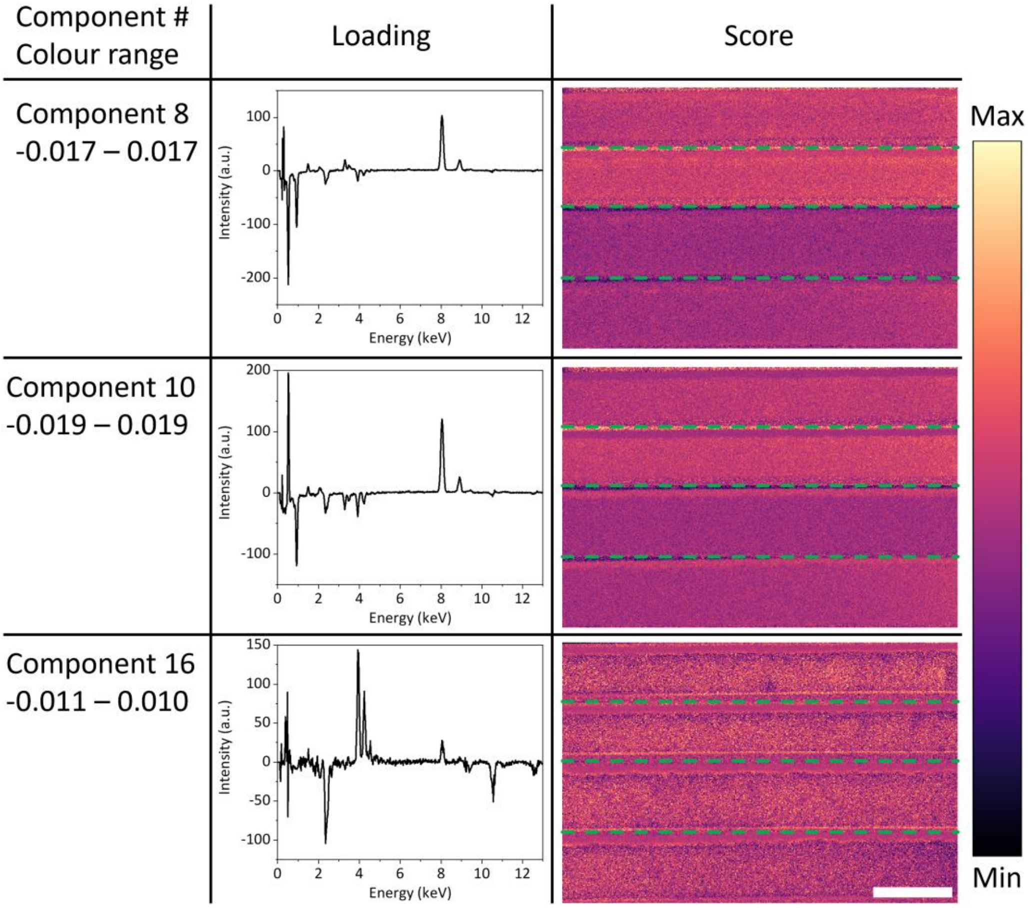


Fig. 4. PCA principal components (8, 10, 16) appearing between the first and second elbow points in the scree plot shown in Figure 2. These components represent genuine physical features in the specimen, but would have been excluded from the denoised model if the first elbow point had been taken as the boundary between principal and noise components. Dashed green lines mark the boundaries between scan areas. Scale bar represents 1 μm.

Construction of Denoised Model from PCA and NMF Components

The previous section has shown that the elbow point method is not necessarily the best way to sort out the principal and noise components in PCA. While this is a useful finding, identification of the principal components is usually not the final goal of data postprocessing. Rather, it is merely an intermediate step required to construct a denoised model of the raw data, on which further analysis can be performed. Therefore, it is important to investigate how the choice of principal components affects the denoised models produced through MVA.

Five denoised spectrum images (SIs) were constructed to model the raw STEM–EDX dataset introduced above. These are named PCA 4, PCA 16, PCA M, NMF 4, and NMF 16. The PCA 4 model was built using principal components identified by the standard elbow point method, namely PCA components 1–4. The PCA 16 model includes 16 PCA components located before the secondary elbow point. Manual selection of principal components was used to assemble the PCA M model, which includes components 1–4, 8, 10, and 16. Finally, the NMF 4 and NMF 16 models were set up with all components produced via NMF run with $p = 4$ and $p = 16$, respectively. In all cases, the components were not subjected to inverse noise scaling before they were used to construct the denoised models. Inverse scaling is not necessary for our purposes as EDX data quantification only requires the relative background-corrected intensities between X-ray peaks of interest, not their absolute counts. However, we note that for applications where the absolute signal counts are needed, the inverse noise scaling step should be performed. Each denoised model was then subjected to peak intensity extraction and Cliff–Lorimer quantification to obtain quantified elemental maps (Cliff & Lorimer, 1975). The relevant X-ray peaks here are Pb–L_α (10.5 keV), I–L_α (3.94 keV), C–K_α (0.27 keV), and N–K_α (0.39 keV).

The chemical maps and distribution profiles are shown in Supplementary Figures 5–7 for Pb, I, and C, and in Figures 7–9 for N, I/Pb ratio, and N/Pb ratio, respectively. The distribution profiles were produced by grouping all pixels in each map into bins based on their value, and then counting the number of pixels in each bin. The bin widths are 1 at% for the elemental maps and 0.05 for the ratio maps. All five models indicate slightly lower concentrations of the inorganic elements Pb (Supplementary Fig. 5) and I (Supplementary Fig. 6) in the aerosol-treated sample. For both elements, the PCA 4 and NMF 4 models have the least amount of noise as expected from their low number of components, with their distribution profiles dropping sharply at the high concentration end. PCA 16 has the noisiest maps as indicated by its wide distribution profiles. PCA M model results in visibly less noisy maps and narrower distribution profiles compared to PCA 16. Importantly, the control sample maps produced from PCA 16 and M show Pb-rich areas which are

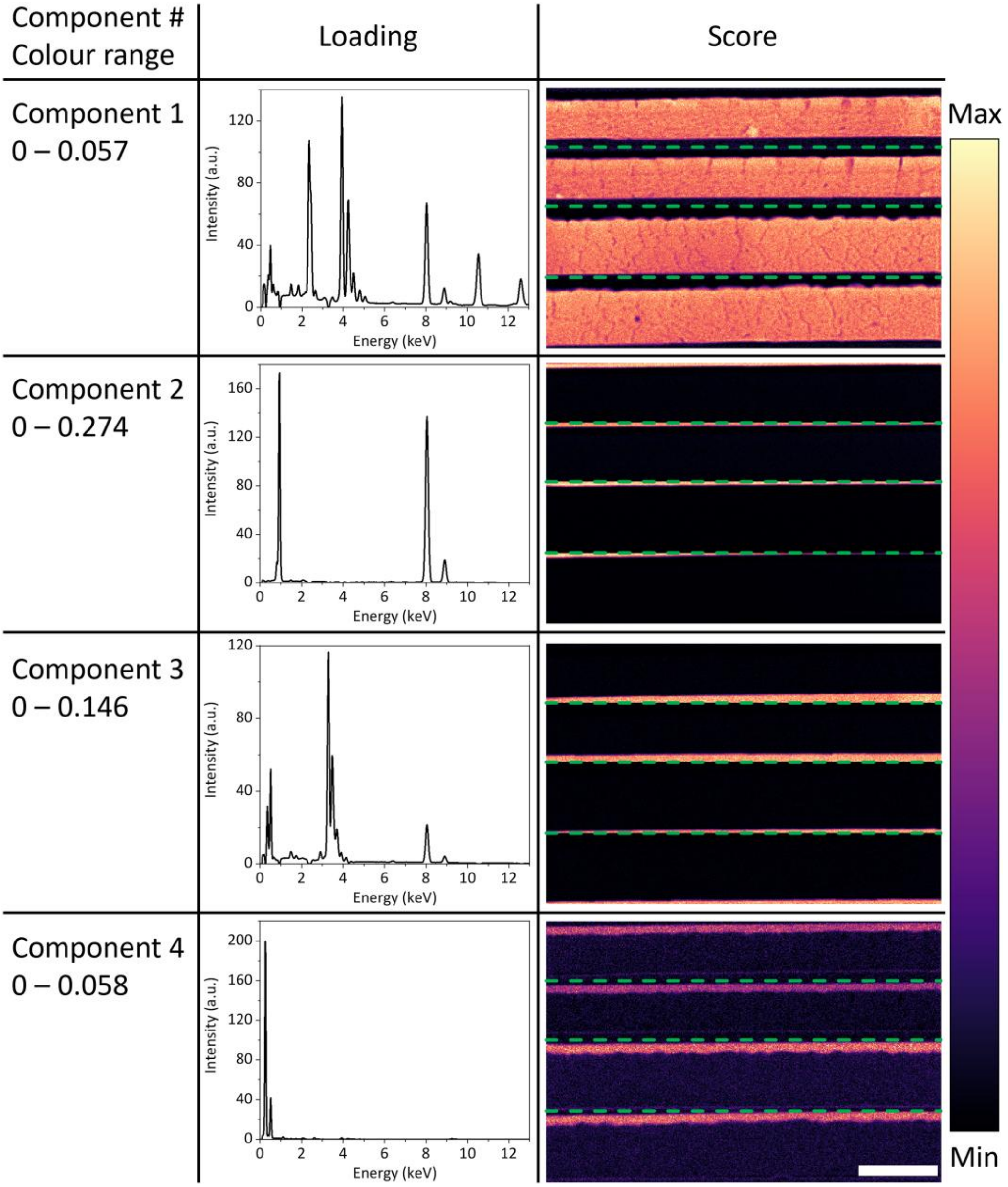


Fig. 5. NMF components when $p = 4$. Dashed green lines mark the boundaries between scan areas. Scale bar represents 1 μm.

attributable to PCA component 16 (Fig. 4). The effect of including this component is most obvious in the I/Pb ratio maps (Fig. 7). The PCA 4 and both NMF models produce very sharp distribution profiles with perfectly uniform perovskite layers where every single pixel has practically the same I/Pb ratio, even where the perovskite was damaged in the bottom scan area. However, this is not an accurate portrayal of the specimens. The PCA 16 and M maps clearly show areas where the local I/Pb ratio is lower. The shape and location of these areas can be matched to features visible in the dark field images (Fig. 1) and in PCA component 16 (Fig. 4), proving that they are real features rather than noise.

As for the organic elements, the C maps and distribution profiles (Supplementary Fig. 7) are largely similar for the five models. The higher C concentration in the aerosol-treated sample is visible in all cases, as is the high C content in the damaged perovskite area, which is expected due to perovskite vaporization and C deposition by prolonged exposure (a few seconds) to the focused electron beam. On the other hand, the N (Fig. 8) and N/Pb ratio (Fig. 9) distributions illustrate the differences between the five denoised models very well. The PCA 4 and both NMF models do not show higher N content and N/Pb ratio in the aerosol-treated sample as PCA 16 and M do. Furthermore, PCA 4, NMF 4, and NMF 16 also suggest that both the N and N/Pb distributions are uniform throughout the perovskite layer. Meanwhile, the PCA 16 and M maps show a thin strip at the top of the perovskite layer (near the perovskite–PCBM interface) where the N concentration and N/Pb ratio is lower than the rest of the perovskite. This is in excellent agreement with the dark field images (Fig. 1) which appear brighter at the same locations, indicating that there are fewer light atoms or more heavy atoms there. These differences are attributable to the inclusion of PCA components 8, 10, and 16 (Fig. 4) in the denoised model. Comparing the PCA 16 and M maps, the high noise in the PCA 16 maps obscures features with low N content to the

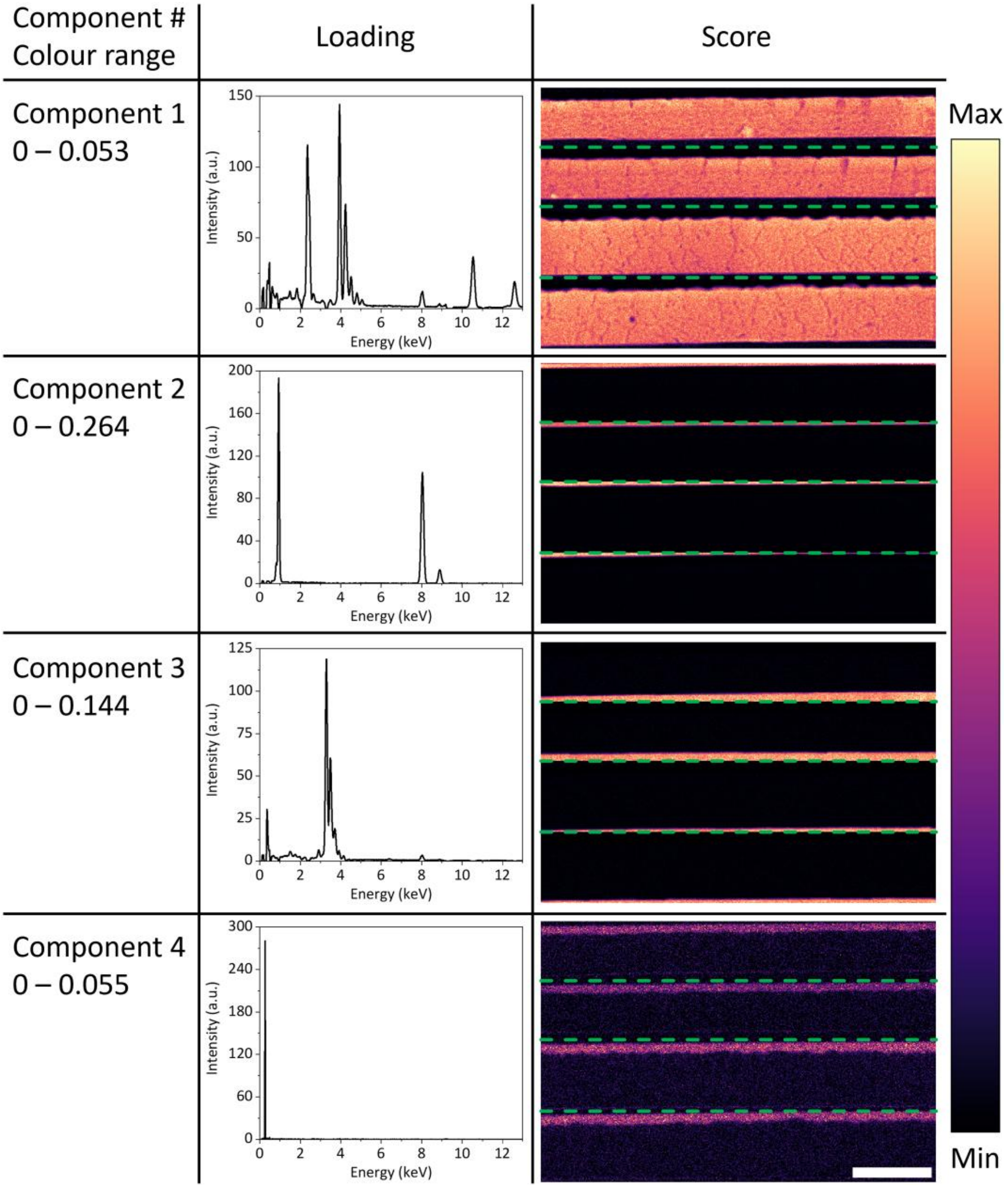


Fig. 6. The first four NMF components when $p = 16$. These components are noticeably dissimilar compared to the $p = 4$ case shown in Figure 5, highlighting the importance of choosing the correct p value. Dashed green lines mark the boundaries between scan areas. Scale bar represents 1 μm.

point that they are hardly visible. PCA M produces the most accurate N and N/Pb maps, correctly showing higher N content in the aerosol-treated sample, lower N content at the perovskite/PCBM interface in the control sample, and the N-poor features distributed in the perovskite layer.

Differences in the accuracy and noise level of elemental maps produced from the five denoised models are summarized in Table 1. Overall, it is concluded that PCA with manual selection of the principal components is the best approach to construct the denoised model of the original dataset. This method accurately reproduces the existence of small heterogeneities in the specimen without including excessive noise.

Selection of PCA Components with Deep Learning

The superiority of manual classification of PCA components in terms of the associated denoised model's accuracy and noise level does not negate the fact that it reintroduces the possibility of operator bias, the absence of which is a strong point in favor of the elbow point method. Furthermore, manual selection can also be very laborious and time-consuming, as PCA can easily produce hundreds of components. To ameliorate these problems, we developed a novel workflow based on a VGGnet-inspired CNN to automate the principal component selection process (Simonyan & Zisserman, 2014).

Two critical challenges in the application of CNNs for electron microscopy data analysis are data insufficiency and class imbalance. The dataset used for image classification tasks needs to contain images and their corresponding class labels. Creating a training dataset with thousands of labeled images is often not feasible in electron microscopy (Siddique et al., 2021). Furthermore, in the context of PCA, there is a far greater number of noise components than principal components. Therefore, the dataset suffers from severe class imbalance, which can lead to deep learning model over-fitting.

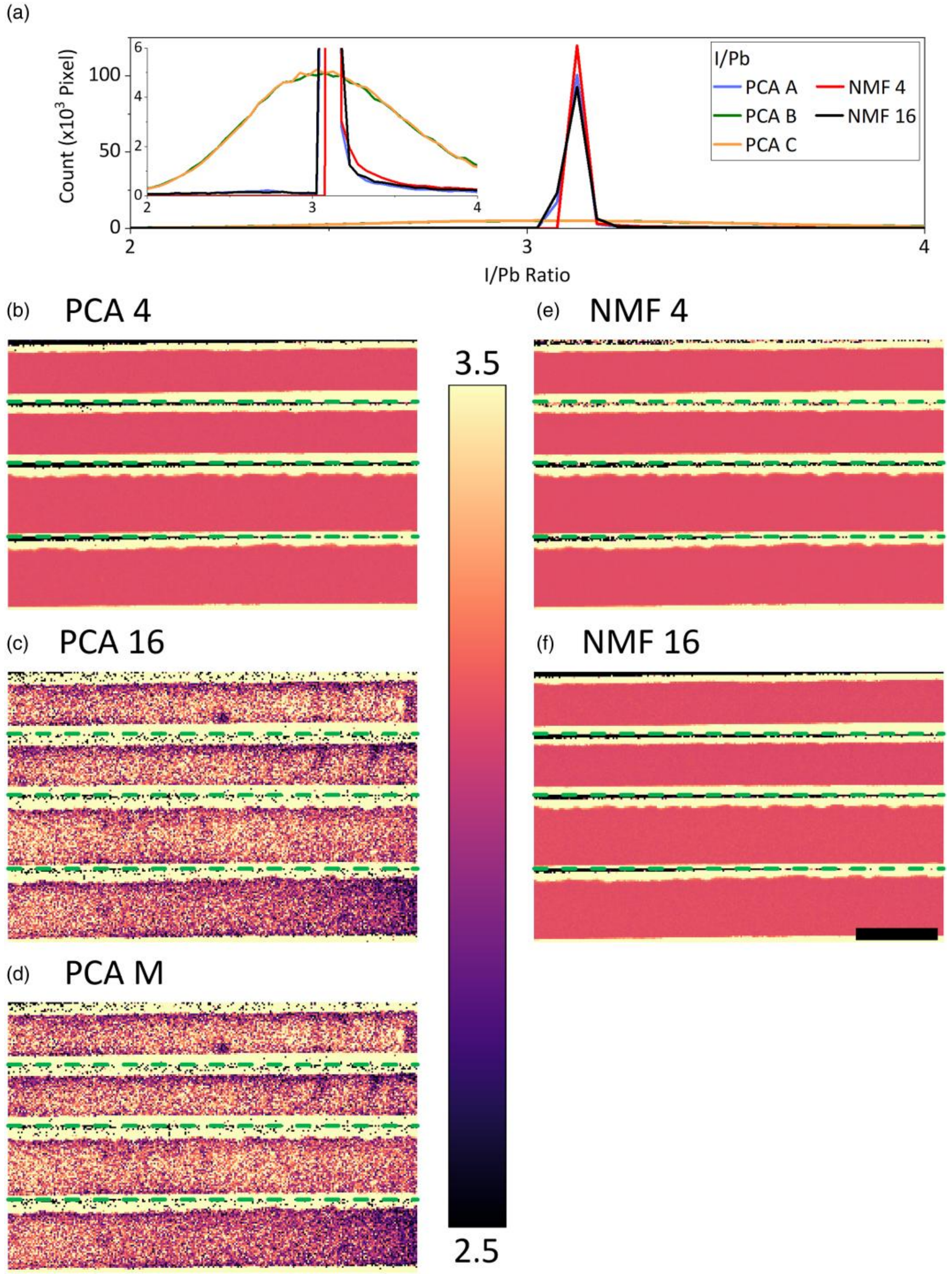


Fig. 7. (**a**) Distributions of I/Pb ratio and (**b–e**) quantified I/Pb ratio maps extracted from the (**b**) PCA 4, (**c**) PCA 16, (**d**) PCA M, (**e**) NMF 4, and (**f**) NMF 16 denoised models. Only the PCA 16 and PCA M models produce accurate ratio maps, while most of the spatial heterogeneity is lost in the other models. The I/Pb ratio maps were spatially rebinned by a factor of 2 to reduce noise. Scale bar is 1 μm.

Our workflow, which was specially designed to overcome the two issues just described, is illustrated in Supplementary Figure 8 and described in detail in the Materials and Methods section. First, we performed PCA on the stack of EDX spectrum images discussed above and produced the scores of PCA components 1–100 in the form of 8-bit RGB (red, green, blue) images. This image format was chosen, as CNNs often make predictions based on color information, especially if the model is trained from scratch, which is the case here (Singh et al., 2020). We manually labeled them as either signal (principal components) or noise (noise components) based on spatial correlation of the features shown in the score images, or lack thereof. Then, we de-stacked each of the scores into four separate images, each corresponding to a different PSC (according to the EDX scan area boundaries shown in Fig. 1) in order to increase the size of our training and testing data. For each of the 400 score images, we normalized the RGB values in the pixels and calculated the means (μ) and standard deviations (σ) of the three RGB color distribution across all pixels. Then, the RGB values in each pixel were

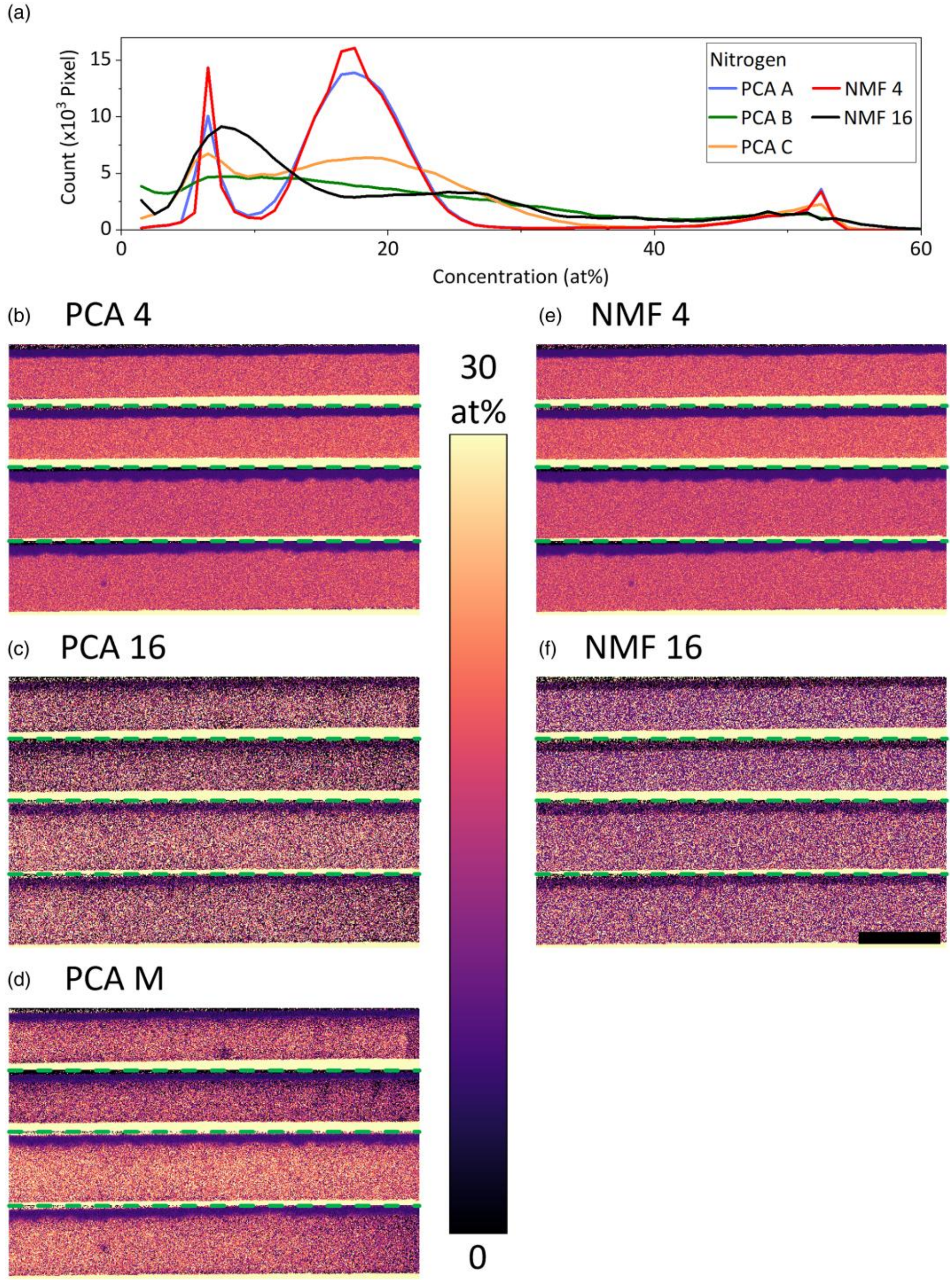


Fig. 8. (**a**) Distribution of N concentration and (**b–e**) N maps extracted from the (**b**) PCA 4, (**c**) PCA 16, (**d**) PCA M, (**e**) NMF 4, and (**f**) NMF 16 denoised models. Only the PCA M model produces an accurate map that shows localized areas with lower N content. Scale bar is 1 μm.

standardized as follows:

$$R_{\text{std}} = \frac{R-\mu_R}{\sigma_R},\ G_{\text{std}} = \frac{G-\mu_G}{\sigma_G},\ B_{\text{std}} = \frac{B-\mu_B}{\sigma_B}$$

This was done to reduce the high numerical variation in the RGB color distributions by narrowing the range of possible values from 0 to 255 to approximately −6 to +8 (Supplementary Fig. 9). Doing so reduces the data complexity and accelerates the model training process. An example of a de-stacked score image and the histogram of its RGB value distribution before and after the normalization and standardization steps is shown in Supplementary Figure 9.

Because the number of noise score images (93) is much higher than the signal ones (7), there is significant class imbalance between the majority (noise) and minority (signal) data classes. To solve this problem, the signal score images are oversampled with offline data augmentation (Afzal et al., 2019). We oversampled the minority class by 14 times, such that it contains approximately the same number of score images as the majority

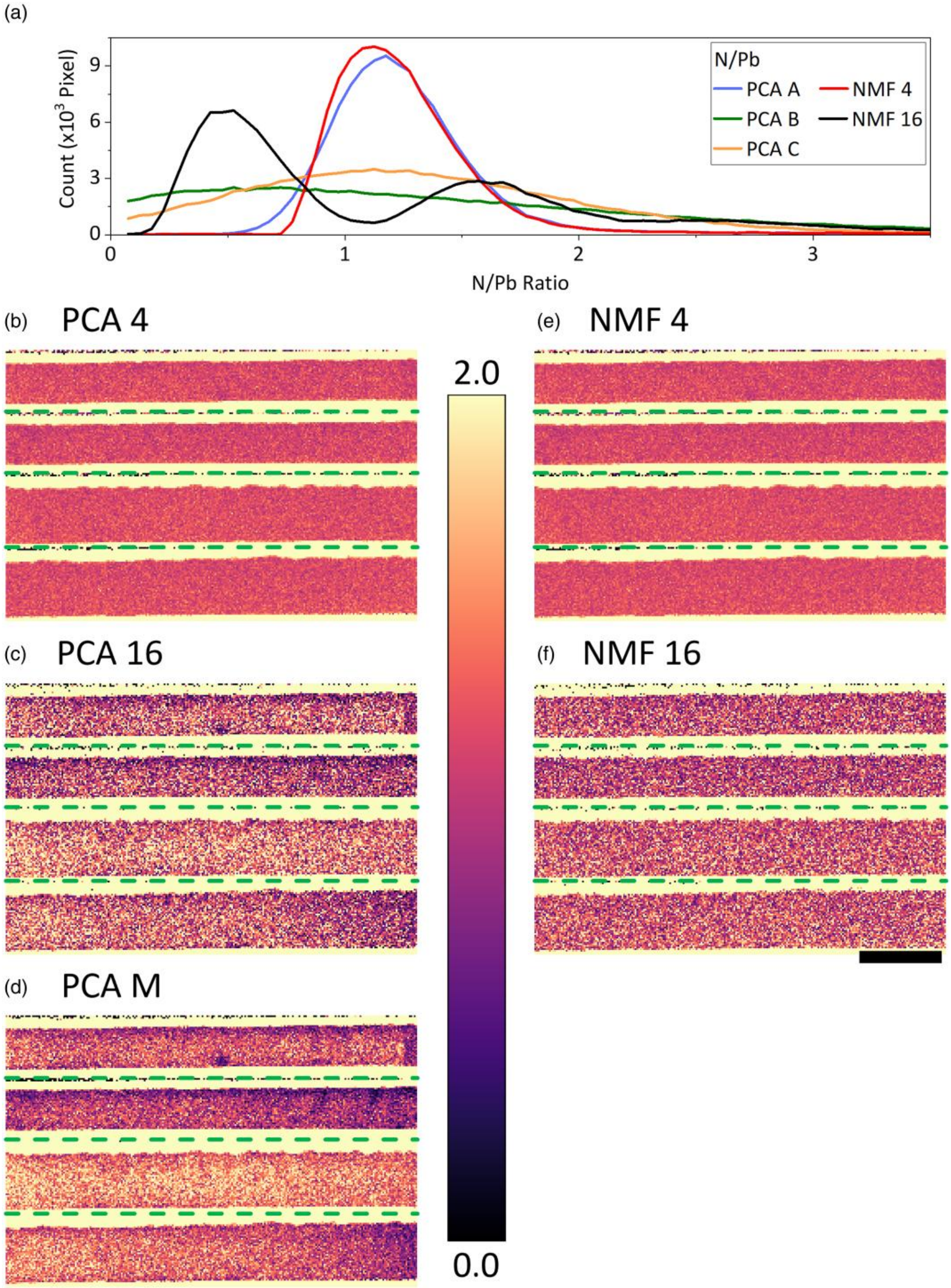


Fig. 9. (**a**) Distribution of N/Pb ratio and (**b–e**) N/Pb ratio maps extracted from the (**b**) PCA 4, (**c**) PCA 16, (**d**) PCA M, (**e**) NMF 4, and (**f**) NMF 16 denoised models. Only the PCA M model produces an accurate map that shows localized areas with lower N/Pb ratio. The N/Pb ratio maps were spatially rebinned by a factor of 2 to reduce noise. Scale bar is 1 μm.

class (see Supplementary Fig. 10). To illustrate the impact of data augmentation, we trained and evaluated our CNN on both the imbalanced and the augmented score image datasets. Table 2 summarizes the CNN evaluation results for the two datasets. To judge the performance of the CNN workflow, we used several parameters as defined in the Materials and Methods section, namely TP fraction, TN fraction, FP fraction, FN fraction, accuracy, sensitivity, specificity, PPV, NPV, and processing time. As shown by the TP and FP fractions, when trained with the imbalanced image dataset, the model is extremely over-fitted and therefore strongly biased toward predicting the majority class only. Due to this bias, the model produced excessively high TN and FN fractions even though it still reached an accuracy of ~95%. In contrast, our CNN model attained an accuracy of >99% when trained with the augmented score images. In addition, the sensitivity value increased dramatically from 0.000 for the imbalanced score images to 0.984 after training the model with the augmented image dataset. This excellent performance suggests that the class imbalance problem has been tackled and the model is no longer

Table 1. Classification of Accuracy (A) and Noise Level (N) of the Quantified Elemental and Ratio Maps Produced from the Five Denoised Models.

Element / Ratio	PCA 4		PCA 16		PCA M		NMF 4		NMF 16	
	A	N	A	N	A	N	A	N	A	N
Pb	🟥	🟩	🟩	🟥	🟩	🟨	🟥	🟩	🟥	🟩
I	🟩	🟩	🟩	🟥	🟩	🟨	🟩	🟩	🟩	🟥
I/Pb	🟥	🟩	🟩	🟨	🟩	🟨	🟥	🟩	🟥	🟩
C	🟩	🟩	🟩	🟨	🟩	🟨	🟩	🟩	🟩	🟨
N	🟥	🟩	🟩	🟥	🟩	🟨	🟥	🟩	🟥	🟥
N/Pb	🟥	🟩	🟩	🟥	🟩	🟨	🟥	🟩	🟥	🟥

A is classified as either accurate (green) or not (red), while N is ranked as low noise (green), medium (yellow), and high noise (red).

biased. The high PPV (1) and NPV (0.95) scores of the CNN model trained on the augmented image dataset indicate that it is a reliable predictor of whether a PCA component should be classified as a principal or a noise component. It is also worth mentioning that the end-to-end deep learning model reached the >99% accuracy value after a short training time of only 46.3 s. The training step only needs to be done once in order to find the weight coefficients of the neural network model, which then can be applied many times more quickly to similar datasets without further training.

Our CNN model outputs not only its prediction (signal or noise) for each PCA component score but also the degree of confidence with which it made that prediction. To illustrate the utility of this feature, three representative PCA scores are shown in Table 3. The first digit in the score code refers to the PCA component index a particular scan area's score image was de-stacked from, while the second digit is that scan area's order in the original score stack, counted from the top. Compared with score 5_1, the neural network is clearly more confident when classifying score 1_1, as expected due to the latter exhibiting much more apparent features. Score 8_2 is an example of an interesting case. The stacked component 8 is clearly a principal component due to the apparent changes in signal intensity from one scan area to the next (see Fig. 4). However, this intensity variation is lost during the de-stacking step in score image preprocessing, so score 8_2 does not possess significant signal variation within its scan area and the CNN correctly classifies it as noise. This strongly suggests that our model is robust and capable of accurate PCA component classification even when data does not come from a number of *a priori* dissimilar specimens.

Table 2. Summary of the Model Evaluation Results When Training the Model with the Imbalanced and Augmented Loading Images.

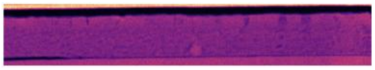
Evaluation Parameter	Imbalanced Dataset	Augmented Dataset
TP fraction	0.0%	55.0%
TN fraction	95.0%	44.1%
FP fraction	0.0%	0.0%
FN fraction	5.0%	0.9%
Accuracy	95.0%	99.1%
Sensitivity	0.000	0.984
Specificity	1	1
PPV	N/A	1
NPV	0.95	0.99
Time taken	27.6 s	46.3 s

TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predicted value; NPV, negative predicted value.

Conclusion

Nanoscale materials' characterization is an invaluable tool in materials science, but it is not without its limitations. One of these is the likelihood of specimen damage in beam-sensitive materials induced by the high-energy and high-intensity probe beams. Therefore, it is crucial that the probe dose is minimized to suppress beam-induced damage and ensure that valid conclusions can be drawn from the characterization data. Dose minimization would also reduce the signal count and hence measurement accuracy, but this disadvantageous effect can be ameliorated using MVA algorithms and CNN. This work reports the optimization and validation of PCA and NMF

Table 3. Examples of PCA Loadings and the Confidence with Which the Trained CNN Model Predicted Their Classification as Signal or Noise.

Code	Score	Prediction Confidence	Predicted Label	Ground Truth
1_1		Noise: 0.4% Signal: 99.6%	Signal	Signal
5_1		Noise: 76.7% Signal: 23.3%	Noise	Noise
8_2		Noise: 80.0% Signal: 20.0%	Noise	Signal

for postprocessing of multidimensional characterization data acquired from heterogeneous beam-sensitive specimens. We conclude that NMF works well for data decomposition into components that are easily recognizable as physical features. However, PCA should be performed first, such that the produced scree plot can be used to determine the number of output NMF components. For data denoising, we find that the oft-used approach of scree plot-based separation between principal and noise components in PCA leads to suboptimal denoised models, even after the Poissonian noise distribution in the original dataset has been scaled. Manual selection of principal components provides a better balance between the necessity of recognizing real physical features and maximizing SNR, but is time-consuming and may enable operator bias. To obtain a fast, automated, and bias-free principal component selection, we present a VGG-inspired five-layer CNN model. Our results indicate that the deep learning-based method recognizes principal components accurately in a short computation time (~2 components/s even when trained from scratch). The principal components identified by CNN can then be used to construct an accurate denoised model of the original data. The combined MVA–CNN workflow presented here can be useful not only for STEM–EDX but also for other nanoscale characterization techniques whose data is amenable to MVA, such as optical spectroscopy, electron energy loss spectroscopy, X-ray absorption spectroscopy, scanning probe microscopy, secondary ion mass spectroscopy, and more (Mak et al., 2014; Kannan et al., 2018; Trindade, 2018). Due to data limitation, we have not tested our model on signals produced from those techniques. However, the trained weights in the neural network can be directly used for transfer-learning with the same network architecture when totally different types of signals are provided. With the new dataset, accurately retraining the network from scratch with the same steps as summarized in this work can be done at a rate as quickly as ~2 components/s. We believe the proposed approach can provide an intuitive way to link artificial intelligence with materials' characterization.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1093/micmic/ozad033.

Acknowledgments

F.U.K. acknowledges the Jardine Foundation and Cambridge Trust for a doctoral scholarship. The authors acknowledges the Horizon 2020 Framework Programme 823717 - ESTEEM3.

Financial Support

T.D., S.R.R., and J.B. thank the EPSRC Plastic Electronics CDT (EP/L016702/1) for financial support and provision of equipment resource.

Conflict of Interest

The authors declare that they have no competing interest.

References

Afzal S, Maqsood M, Nazir F, Khan U, Aadil F, Awan K, Mehmood I & Song O (2019). A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. *IEEE Access* 7, 115528–115539. https://doi.org/10.1109/ACCESS.2019.2932786

Azpiroz JM, Mosconi E, Bisquert J & De Angelis F (2015). Defect migration in methylammonium lead iodide and its role in perovskite solar cell operation. *Energy Environ Sci* **8**, 2118–2127. https://doi.org/10.1039/C5EE01265A

Burnett TL & Withers PJ (2019). Completing the picture through correlative characterization. *Nat Mater* **18**, 1041–1049. https://doi.org/10.1038/s41563-019-0402-8

Cattell RB (1966). The scree test for the number of factors. *Multivar Behav Res* **1**, 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chen Q, Dwyer C, Sheng G, Zhu C, Li X, Zheng C & Zhu Y (2020). Imaging beam-sensitive materials by electron microscopy. *Adv Mater* **32**, 1907619. https://doi.org/10.1002/adma.201907619

Chen Y & Zhou H (2020). Defects chemistry in high-efficiency and stable perovskite solar cells. *J Appl Phys* **128**, 060903. https://doi.org/10.1063/5.0012384

Cliff G & Lorimer GW (1975). The quantitative analysis of thin specimens. *J Microsc* **103**, 203–207. https://doi.org/10.1111/j.1365-2818.1975.tb03895.x

de la Peña F, Prestat E, Fauske VT, Burdet P, Furnival T, Jokubauskas P, Nord M, Ostasevicius T, MacArthur KE, Johnstone DN, Sarahan M, Lahnemann J, Taillon J, Aarholt T, Migunov V, Eljarrat A, Caron J, Mazzucco S, Martineau B, Somnath S, Poon T, Walls M, Slater T, Tappy N, Cautaerts N, Winkler F, Donval G & Myers JC (2020). HyperSpy 1.6.1. https://doi.org/10.5281/zenodo.7263263

Doherty TAS, Winchester AJ, Macpherson S, Johnstone DN, Pareek V, Tennyson EM, Kosar S, Kosasih FU, Anaya M, Abdi-Jalebi M, Andaji-Garmaroudi Z, Wong EL, Madéo J, Chiang Y-H, Park J-S, Jung Y-K, Petoukhoff CE, Divitini G, Man MKL, Ducati C, Walsh A, Midgley PA, Dani KM & Stranks SD (2020). Performance-limiting nanoscale trap clusters at grain junctions in halide perovskites. *Nature* **580**(7803), 360–366. https://doi.org/10.1038/s41586-020-2184-1

Du T, Burgess CH, Kim J, Zhang J, Durrant JR & McLachlan MA (2017). Formation, location and beneficial role of PbI_2 in lead halide perovskite solar cells. *Sustainable Energy Fuels* **1**, 119–126. https://doi.org/10.1039/C6SE00029K

Du T, Ratnasingham SR, Kosasih FU, Macdonald TJ, Mohan L, Augurio A, Ahli H, Lin C, Xu S, Xu W, Binions R, Ducati C, Durrant JR, Briscoe J & McLachlan MA (2021). Aerosol assisted solvent treatment: A universal method for performance and stability enhancements in perovskite solar cells. *Adv Energy Mater* **11**(33), 2101420. https://doi.org/10.1002/aenm.202101420

Eames C, Frost JM, Barnes PRF, O'Regan BC, Walsh A & Islam MS (2015). Ionic transport in hybrid lead iodide perovskite solar cells. *Nat Commun* **6**, 7497. https://doi.org/10.1038/ncomms8497

Haruyama J, Sodeyama K, Han L & Tateyama Y (2015). First-principles study of ion diffusion in perovskite solar cell sensitizers. *J Am Chem Soc* **137**, 10048–10051. https://doi.org/10.1021/jacs.5b03615

Henderson R (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28**, 171–193. https://doi.org/10.1017/S003358350000305X

HyperSpy (n.d.*a*). hyperspy.learn.svd_pca module. Available at http://hyperspy.org/hyperspy-doc/current/api/hyperspy.learn.svd_pca.html#hyperspy.learn.svd_pca.svd_pca (retrieved May 16, 2022).

HyperSpy (n.d.*b*). hyperspy.learn.mva module. Available at http://hyperspy.org/hyperspy-doc/current/api/hyperspy.learn.mva.html#hyperspy.learn.mva.MVA.decomposition (retrieved May 16, 2022).

Ilett M, S'ari M, Freeman H, Aslam Z, Koniuch N, Afzali M, Cattle J, Hooley R, Roncal-Herrero T, Collins SM, Hondow N, Brown A & Brydson R (2020). Analysis of complex, beam-sensitive materials by transmission electron microscopy and associated techniques.

Philos Trans R Soc A Math Phys Eng Sci **378**, 20190601. https://doi.org/10.1098/rsta.2019.0601

Kannan R, Ievlev AV, Laanait N, Ziatdinov MA, Vasudevan RK, Jesse S & Kalinin SV (2018). Deep data analysis via physically constrained linear unmixing: Universal framework, domain examples, and a community-wide platform. *Adv Struct Chem Imaging* **4**, 6. https://doi.org/10.1186/s40679-018-0055-8

Keenan MR & Kotula PG (2004). Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf Interface Anal* **36**, 203–212. https://doi.org/10.1002/sia.1657

Kosasih FU, Cacovich S, Divitini G & Ducati C (2020). Nanometric chemical analysis of beam-sensitive materials: A case study of STEM-EDX on Perovskite solar cells. *Small Methods* **5**(2), 2000835. https://doi.org/10.1002/smtd.202000835

Kosasih FU, Divitini G, Orri JF, Tennyson EM, Kusch G, Oliver RA, Stranks SD & Ducati C (2022). Optical emission from focused ion beam milled halide perovskite device cross-sections. *Microsc Res Tech* **85**(6), 2351–2355. https://doi.org/10.1002/jemt.24069

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

Mak R, Lerotic M, Fleckenstein H, Vogt S, Wild SM, Leyffer S, Sheynkin Y & Jacobsen C (2014). Non-negative matrix analysis for effective feature extraction in X-ray spectromicroscopy. *Faraday Discuss* **171**, 357–371. https://doi.org/10.1039/C4FD00023D

Mohan S, Manzorro R, Vincent JL, Tang B, Sheth DY, Simoncelli EP, Matteson DS, Crozier PA & Fernandez-Granda C (2020). Deep denoising for scientific discovery: a case study in electron microscopy. *IEEE Trans Comput Imaging* **8**, 585–597. https://doi.org/10.1109/TCI.2022.3176536

Paatero P & Tapper U (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126. https://doi.org/10.1002/env.3170050203

Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Lond Edinb Dublin Philos Mag J Sci* **2**, 559–572. https://doi.org/10.1080/14786440109462720

Saliba M, Matsui T, Seo J-Y, Domanski K, Correa-Baena J-P, Nazeeruddin MK, Zakeeruddin SM, Tress W, Abate A, Hagfeldt A & Grätzel M (2016). Cesium-containing triple cation perovskite solar cells: Improved stability, reproducibility and high efficiency. *Energy Environ Sci* **9**, 1989–1997. https://doi.org/10.1039/C5EE03874J

Schlossmacher P, Klenov DO, Freitag B, von Harrach S & Steinbach A (2010). Nanoscale chemical compositional analysis with an innovative S/TEM-EDX system. *Microsc Anal* **24**, S5–S8. https://analyticalscience.wiley.com/do/10.1002/micro.504/full/i2552fe95211c69c3cfa01cd212a664c3.pdf

Scikit Learn (n.d.). sklearn.decomposition.NMF. Available at https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html#sklearn.decomposition.NMF (retrieved May 16, 2022).

Siddique N, Paheding S, Elkin CP & Devabhaktuni V (2021). U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* **9**, 82031–82057. https://doi.org/10.1109/ACCESS.2021.3086020

Simonyan K & Zisserman A (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*. https://doi.org/10.48550/arXiv.1409.1556

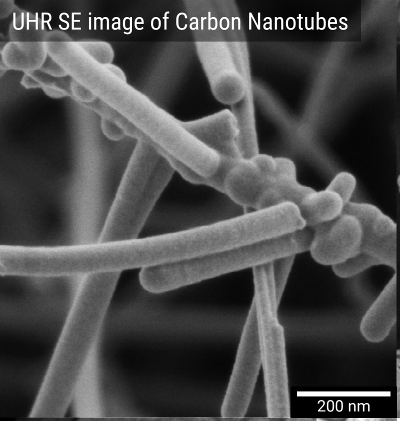
Singh A, Bay A & Mirabile A (2020). Assessing the importance of colours for CNNs in object recognition. *arXiv*. https://doi.org/10.48550/arXiv.2012.06917

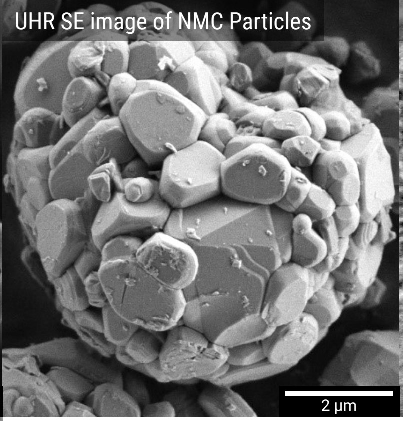
Tennyson EM, Doherty TAS & Stranks SD (2019). Heterogeneity at multiple length scales in halide perovskite semiconductors. *Nat Rev Mater* **4**, 573–587. https://doi.org/10.1038/s41578-019-0125-0

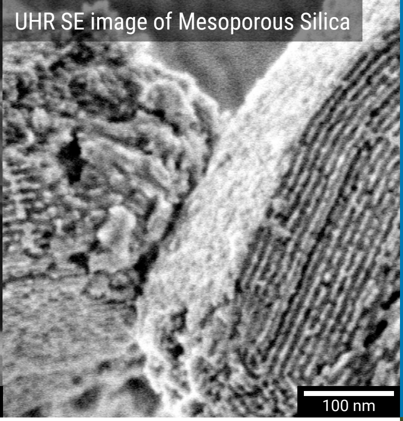
Trindade GF (2018). *The Development of Multivariate Analysis Methodologies for Complex ToF-SIMS Datasets: Applications to Materials Science*. Guildford: University of Surrey.

Zhang C, Han R, Zhang AR & Voyles PM (2020). Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy* **219**, 113123. https://doi.org/10.1016/j.ultramic.2020.113123

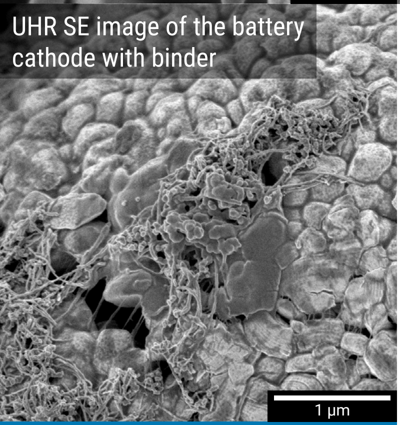
Zhu M & Ghodsi A (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput Stat Data Anal* **51**, 918–930. https://doi.org/10.1016/j.csda.2005.09.010


UHR SE image of Carbon Nanotubes


UHR SE image of NMC Particles


UHR SE image of Mesoporous Silica

Capture surface details at the nanoscale from any materials


UHR SE image of the battery cathode with binder

NEW GENERATION

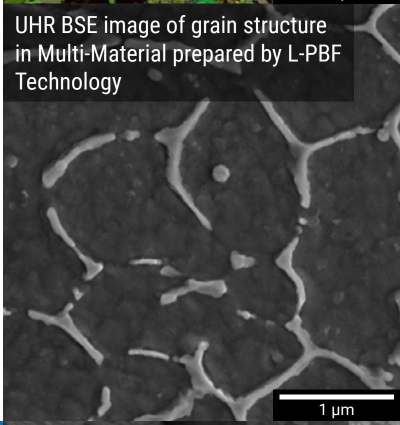
TESCAN CLARA

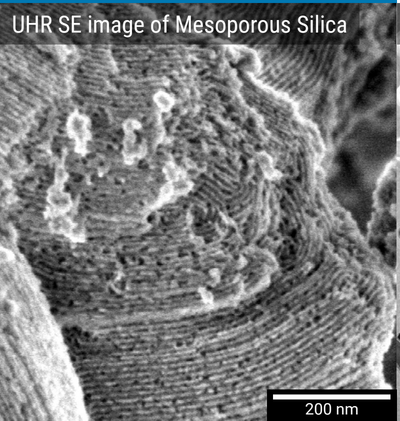
UHR SEM for quick, accurate and comprehensive nanoscale surface analysis of any material

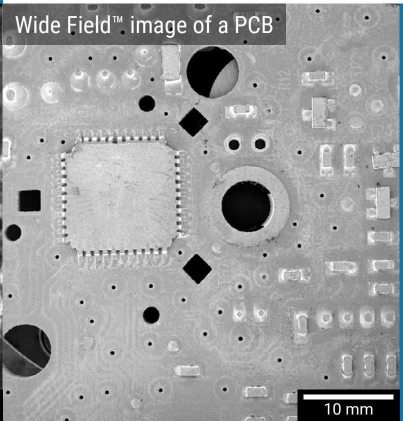
TESCAN

EDS map of the ancient plaster

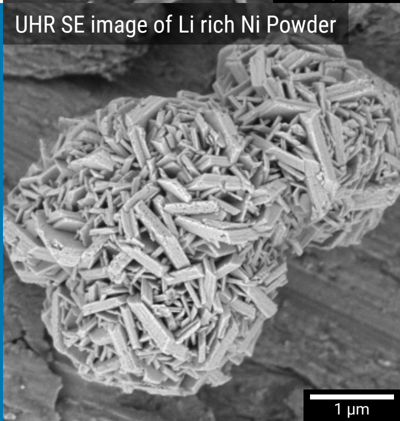
Reveal new contrast information and hidden features


UHR BSE image of grain structure in Multi-Material prepared by L-PBF Technology


UHR SE image of Mesoporous Silica


Wide Field™ image of a PCB

Obtain the right data in a short time disregard your SEM experience


UHR SE image of Li rich Ni Powder