Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3610

Unsupervised multi-clustering and decision-making strategies for 4D-STEM orientation mapping

Junhao Cao,[ab] Nicolas Folastre,[ab] Gozde Oney,[c] Edgar Rauch,[d] Stavros Nicolopoulos,[e] Partha Pratim Das [e] and Arnaud Demortière [*abf]

This study presents a novel integration of unsupervised learning and decision-making strategies for the advanced analysis of 4D-STEM datasets, with a focus on non-negative matrix factorization (NMF) as the primary clustering method. Our approach introduces a systematic framework to determine the optimal number of components (k) required for robust and interpretable orientation mapping. By leveraging the K-component loss method and Image Quality Assessment (IQA) metrics, we effectively balance reconstruction fidelity and model complexity. Additionally, we highlight the critical role of dataset preprocessing in improving clustering stability and accuracy. Furthermore, our spatial weight matrix analysis provides insights into overlapping regions within the dataset by employing threshold-based visualization, facilitating a detailed understanding of cluster interactions. The results demonstrate the potential of combining NMF with advanced IQA metrics and preprocessing techniques for reliable orientation mapping and structural analysis in 4D-STEM datasets, paving the way for future applications in multi-dimensional material characterization.

Received 20th February 2025
Accepted 15th October 2025

DOI: 10.1039/d5dd00071h

rsc.li/digitaldiscovery

Introduction

Recent advancements in scientific instruments for material analysis have led to the development of devices able to generate vast amounts of data across multiple modalities with high spatial resolution. These large and complex datasets often require advanced AI algorithms for efficient processing. For instance, in transmission electron microscopy (TEM) field, the recent integration of advanced techniques such as hybrid-pixel detection, electron beam precession and highly coherent beam has culminated in the emergence of a new class of hyperspectral analysis known as four-dimensional scanning transmission electron microscopy (4D-STEM).[1a] The 4D-STEM technique involves acquiring a two-dimensional dataset of diffraction patterns over a two-dimensional scanning region, resulting in a four-dimensional dataset.[1b] A single 4D-STEM dataset can contain more than 100k electron diffraction patterns (512×512 px^2). Consequently, storing, managing, and efficiently analyzing such data presents significant challenges. The complexity of 4D-STEM data stems from the multidimensional nature of the structural information encoded within each diffraction pattern,[24] involving advanced algorithms and significant computational resources for effective analysis. Moreover, pre-processing electron diffraction patterns to maintain data integrity and mitigate various noise sources introduces significant computational overhead and remains a time-consuming task. Consequently, automated data processing leveraging advanced algorithms, particularly those integrating statistical methods and machine learning techniques, is crucial for enhancing efficiency and accuracy in diffraction pattern analysis.

Pattern matching strategies based on pixel-to-pixel cross-correlation coefficients between experimental patterns and simulated patterns, generated from known crystallographic structure data in Crystallographic Information File (CIF) format,[2a] have been extensively employed for the analysis of 4D-STEM datasets. This method, which facilitates the extraction of orientation and phase maps, has been implemented in several software packages, including Astar,[2b] py4D-STEM,[2c] and pyXEM.[2d] The automated crystal orientation mapping (ACOM) procedure determines the orientation of each diffraction pattern, enabling accurate crystallographic analysis of materials. However, electron diffraction patterns are inherently sparse datasets, with fewer than 10% of the pixels containing meaningful signals. Thus, the implementation of data reduction strategies, which convert sparse data into dense representations, can significantly enhance post-processing

[a]*Laboratoire de Réactivité et de Chimie des Solides (LRCS), CNRS UMR 7314, Université de Picardie Jules Verne, Hub de l'Energie, Rue Baudelocque, 80039 Amiens Cedex, France. E-mail: arnaud.demortiere@cnrs.fr*
[b]*Réseau sur le Stockage Electrochimique de l'Energie (RS2E), CNRS FR 3459, Hub de l'Energie, Rue Baudelocque, 80039 Amiens Cedex, France*
[c]*Institut de Chimie de la Matière Condensée de Bordeaux (ICMCB), Bordeaux, France*
[d]*Université Grenoble Alpes, CNRS, Grenoble INP, SIMAP, 38000 Grenoble, France*
[e]*NanoMegas Company, Belgium*
[f]*ALISTORE-European Research Institute, CNRS FR 3104, Hub de l'Energie, Rue Baudelocque, 80039 Amiens Cedex, France*

efficiency for feature extraction, clustering, and reconstruction, as demonstrated in the development of ePattern (see in SI Algorithm SI_6).[42]

Clustering and data reduction strategies are standard techniques for handling large and high-dimensional datasets. Their objective is to enhance data interpretability while preserving the most relevant information from the original dataset.[3,4] For instance, Principal Component Analysis (PCA) is a widely used unsupervised learning technique for dimensionality reduction, transforming data into a new coordinate system to capture most of the variance in fewer dimensions.[4–6] While effective in applications like image processing, noise reduction, and data compression, PCA has limitations, including its inability to capture non-linear data structures and the interpretability challenges posed by negative component values.[11,12] Furthermore, when combined with clustering algorithms, PCA's results can be sensitive to the user-defined number of clusters, potentially affecting analysis robustness.[13]

In contrast, Non-negative Matrix Factorization (NMF)[10a] offers several advantages over PCA in the context of unsupervised learning and data dimensionality reduction. Unlike PCA, which allows for both positive and negative components, NMF imposes non-negativity constraints on the factorized matrices. This non-negativity constraint results in a parts-based data representation, making NMF highly effective for interpreting and extracting meaningful features in applications such as image processing, text mining, and spectral data analysis. Furthermore, NMF is better suited for handling non-linear and non-convex data structures compared to the PCA. While both PCA and NMF are linear factorization methods, NMF's non-negativity constraints and additive structure enable it to better approximate non-linear and non-convex data patterns common in practical applications. This makes NMF more suitable for tasks where data is composed of localized, interpretable parts, even if the overall manifold is non-linear. The additive nature of NMF components can capture the underlying data patterns more effectively when the data consists of overlapping or additive features. NMF's powerful ability to extract subtle orientation variations has been utilized to enhance the accuracy and reliability of detecting different crystal orientations in 4D-STEM datasets.[27]

In traditional clustering methods, the determination of the optimal number of clusters is inherently challenging due to several factors.[27b] The intrinsic complexity of the data can make the natural separations between clusters unclear, especially in the presence of overlapping clusters, noise, or varying density and shape.[10b] The absence of ground truth in many clustering applications requires reliance on data-driven methods to estimate the optimal number of clusters.[28] To tackle these challenges, various methods have been proposed. The elbow method entails plotting the within-cluster sum of squares (WCSS) against the number of clusters to identify a point where adding more clusters yields diminishing returns.[10c] Silhouette analysis assesses cluster compactness and separation, selecting the number of clusters that maximizes the silhouette score.[10d] Incorporating domain knowledge can also guide and validate the clustering process, ensuring alignment with practical expectations.[10e] By integrating these approaches and validating results across multiple criteria, the determination of the optimal number of components in clustering becomes more robust and reliable.

Brute-force or sophisticated methods for determining the optimal number of clusters usually involves running the clustering algorithm multiple times, each with a different number of clusters, and selecting the configuration that yields the most favorable results. These approaches are computationally intensive. To address this issue more effectively, integrating decision-making approaches, such as multi-criteria decision-making techniques, can provide substantial advantages by automating the selection process and enhancing the robustness of the clustering outcomes. Decision-making can be considered as a problem-solving method providing an optimal solution to a specific event.[14,15] After analyzing of a finite set of alternative solutions, the objective is to categorize these alternatives to establish a priority ranking among them. Generally, the conception of decision-making in unsupervised learning[17] is related to extracting significant patterns, features, or underlying information,[16] without specific labels, revealing the inherent characteristics or relationships hidden in the raw data.[18,19] In the 4D-STEM data clustering process, decision-making involves several considerations specific to the qualities and attributes of electron diffraction pattern datasets, which encompass both crystal orientation and crystallographic phase information.[20]

An additional significant challenge in 4D-STEM mapping is the overlap of patterns from different crystals.[20a,b] In 4D-STEM, diffraction patterns are generated from probe positions scanning crystals that may be in proximity or/and superimposed configurations.[2,20] Thus, assigning the correct crystallographic orientation becomes difficult when overlapping occurs.[25] The diffraction patterns in 4D-STEM can demonstrate complicated features and overlapping spots, the ambiguity of which leads to requiring accurate interpreting of the orientation.[24] The complexity of diffraction patterns can cause errors or uncertainties regarding the determining crystal orientations.[26,27] Efficient algorithms for overlap detection are thus required to specify the precise location of each individual diffraction pattern.[25a,b]

In this study, we develop clustering approach using Non-negative Matrix Factorization (NMF) to analyze four-dimensional scanning transmission electron microscopy (4D-STEM) datasets for orientation mapping. We introduce an efficient method termed "*K*-component loss," which, when combined with Image Quality Assessment (IQA), enables the automatic and effective detection of material characteristics and clustering within large datasets. Our methodology begins with an evaluation phase (level one) to determine initial NMF parameters. Then, we employ a *k*-metric derived from IQA to ascertain the optimal number of clusters (k) in a subsequent phase (level two). This approach is particularly advantageous for processing overlapping diffraction patterns, as it leverages advanced data analysis techniques to separate overlapping signals, assess the similarity of each component, and accurately extract pertinent features from the dataset. By integrating NMF with IQA, our making-decision method offers a robust framework for the analysis of complex 4D-STEM data, facilitating enhanced material characterization and more precise orientation mapping.

Methods

Non-Negative Matrix Factorization (NMF) algorithm

Non-negative matrix factorization (NMF)[1,2] is a common unsupervised machine learning algorithm that decomposes an original non-negative matrix V into two non-negative matrices, W and H. Popularized by Lee and Seung in 1999,[29] NMF was initially applied in image processing to achieve parts-based representations of face images by combining learned features. Since its introduction, non-negative matrix factorization (NMF) has become a powerful unsupervised learning algorithm, particularly valued for its superior interpretability in uncovering latent features. By decomposing data into non-negative components, NMF facilitates the identification of meaningful patterns, enhancing the understanding of underlying structures in various datasets (Fig. 1).

In the latent space, essential features of the original matrix are extracted by selecting components (denoted as k) whose number is significantly less than the rank of the original matrix V ($k \ll \min(W, H)$). The matrix $V \approx W \times H$ is factorized into two relatively small matrices (W, H) compared with V (original), the dimensionality of these two matrices is $W \times k$ and $k \times H$, respectively.[30] The linear combination of W and H generates an approximated matrix $V' = W \times H$. W matrix can be interpreted as the feature matrix, in which the k-column represents the most k-relevant feature from the original matrix V.[31] H can be interpreted as the coefficient matrix, in which the element is the weight associated with the W matrix. Moreover, the aim of obtaining the result of approximate matrix V' is achieved by minimizing a loss function.[29]

Lee and Seung introduced an alternating optimization method for NMF.[29] Starting with random non-negative initializations of matrices W and H, the algorithm iteratively, Alternating Least-Square (ALS), minimizes the loss function $\|V - WH\|$ using multiplicative update rules. In each iteration, H is updated while keeping W fixed, followed by updating W with H fixed, ensuring that both matrices remain non-negative throughout the process. This procedure continues until the difference between V and its approximation WH falls below a predefined threshold.[30]

Data preparation

Non-negative Matrix Factorization necessitates a two-dimensional ($M \times N$) non-negative input matrix. Given that 4D-STEM datasets are inherently four-dimensional, with dimensions (M, N, x, y), in which (M, N) represent probe positions and (x, y) correspond to pixels within each diffraction pattern (512 × 512 px^2), it is imperative to preprocess these datasets appropriately. Typically, the product $M \times N$ correlates with the dataset's size.[1] Therefore, converting the 4D-STEM dataset into a two-dimensional matrix V is essential for subsequent matrix computations (details in SI).

To ensure dataset integrity following Non-negative Matrix Factorization (NMF), it is essential to assess information loss between the original and factorized matrices. Incorporating L1 regularization,[32] commonly utilized in machine learning to enhance model sparsity, can effectively select pertinent

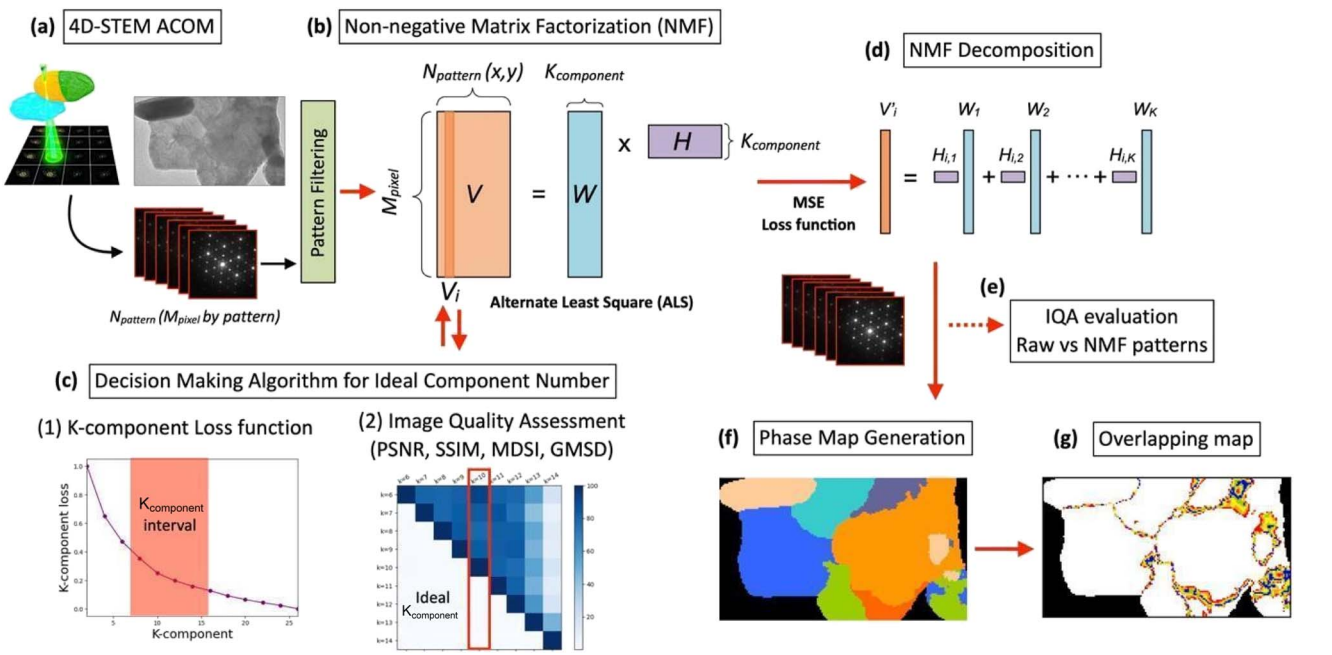

Fig. 1 Schematic representation of the global workflow for clustering and decision-making strategies in the analysis of the 4D-STEM dataset. (a) Overview of the 4D-STEM ACOM acquisition methodology. (b) Hypermatrix decomposition of the original dataset into two matrices, *WWW* and *HHH*, using the Non-negative Matrix Factorization (NMF) algorithm. (c) Core decision-making framework for determining the optimal number of clusters (components): (c_1) initial step identifying the potential range of cluster numbers, and (c_2) similarity assessment of pattern pairs to maximize differentiation while avoiding overfitting. (d) Reconstruction of diffraction patterns based on the NMF-derived results. (e) Image Quality Assessment (IQA) comparing raw diffraction patterns to those reconstructed *via* NMF. (f) Visualization of individual clusters (components) within the dataset, aligned with the optimal component count. (g) Creation of an overlapping map highlighting the regions of cluster co-occurrence.

features, particularly in high-dimensional datasets.[33] This study calculates the difference between the original and factorized matrices, resulting in a K-component loss matrix. We then compute the mean of the absolute values of its elements to quantify information loss.

Noise standard deviation (NSD) to quantify noise influence in diffraction patterns

Noise Standard Deviation (NSD) is a statistical metric used to quantify the magnitude of noise in diffraction patterns, enabling researchers to evaluate how noise impacts data quality, resolution, and interpretability. In diffraction experiments (*e.g.*, X-ray, electron, or neutron diffraction), noise arises from various sources such as detector imperfections, shot noise, thermal fluctuations, and environmental interference. NSD provides a standardized way to characterize this noise, aiding in experimental optimization, algorithm validation, and error analysis. The definition of formula NSD shown in Algorithm SI_1, NSD serves as a metric to quantify the noise level within an image, reflecting the variation or dispersion of pixel values induced by noise.[43] Specifically, it measures the extent to which pixel values deviate from their mean due to noise interference, with higher NSD values indicating more significant noise and lower NSD values suggesting reduced noise. Consequently, by preprocessing the dataset to minimize NSD, NMF is better equipped to focus on extracting meaningful patterns from cleaner, filtered data.[44]

K-Component loss and image quality assessment (IQA)

NMF is a powerful unsupervised learning technique for decomposing high-dimensional data into interpretable basis and coefficient matrices. However, evaluating the quality of the reconstructed dataset V' (derived from NMF) against the original dataset V requires a loss function that balances pixel-wise accuracy and component-wise fidelity. To address this, we propose the K-component loss function as an extension of the Mean Absolute Error (MAE) tailored for NMF-based reconstruction tasks. The K-component loss function is defined as:

$$L_{K\text{-component}} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| V(i,j) - V'(i,j) \right|$$

where V is the original non-negative 4D-STEM dataset. V' is the reconstructed dataset obtained from the NMF model, *i.e.*, $V' = W \times H$, where W and H are the basis and coefficient matrices, respectively. The representation of M and N references to Data preparation and SI. By decomposing the error across $M \times N$, this loss function explicitly quantifies discrepancies in both localized regions and feature-specific patterns, ensuring the reconstruction preserves critical structures inherent to the original data. The parameter K (number of components in NMF) directly influences the loss dynamics. Smaller K simplifies the model but risks under-fitting by over-smoothing fine-grained details, while larger K enhances reconstruction fidelity but may overfit noise or outliers. This trade-off aligns with the K-component loss's ability to guide model selection by highlighting regions or patterns where reconstruction quality diverges significantly from the original data.

According to the K-component loss, the declining trend reflects the loss variation between the NMF results and the original dataset, serving as a reference for evaluating dataset quality. As shown in Fig. 2, the curve is flatter after $k = 10$, indicating that the NMF reaches its performance limit, beyond which further processing offers minimal benefit. Thus, $k = 10$ is identified as a preliminary choice for the number of components. However, to ensure this selection does not lead to overfitting, a secondary evaluation using Image Quality Assessment (IQA) is conducted. For holistic evaluation, the K-component loss can be integrated with perceptual metrics. This combination ensures optimization aligns not only with pixel-wise accuracy but also with human-interpretable quality, making it particularly effective for applications like image denoising, hyperspectral unmixing, or document topic modeling.

IQA objectively analyzes and quantifies image quality through algorithms that estimate perceptual quality based on various features.[34] Its goal is to provide mathematical metrics aligned with human visual perception.[35,36] IQA facilitates the evaluation of image quality, performance analysis of image processing algorithms, and supports decision-making for quality enhancement.[37] IQA methods are generally categorized into Full-Reference (FR) and No-Reference (NR) approaches.[38] FR-IQA, being more established, is commonly used in machine learning for image quality evaluation. It compares a reference (original) image with a target (processed or distorted) image using metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Squared Error (MSE), which are widely applied in FR-IQA.[36]

To quantify the fidelity and difference between each two-diffraction pattern of NMF-reconstructed component maps, we computed four full-reference IQA indices to determine the maximum different orientation in material, where Structural Similarity Index (SSIM) measures perceptual similarity in luminance, contrast and structure. SSIM $\in [-1, 1]$, where 1.00 denotes perfect structural agreement. Values $\geq$0.90 are considered excellent, 0.80–0.90 good, and close to -1 indicate notable dissimilarity. Peak Signal-to-Noise Ratio (PSNR) reflects the ratio between the maximum possible pixel intensity and the mean squared error. Expressed in decibels (dB). PSNR $\geq$ 30 dB generally signifies high-fidelity reconstructions; PSNR < 25 dB suggests significant loss, which usually provides a global view to evaluate the contribution of each clustering. Gradient Magnitude Similarity Deviation (GMSD) assesses local gradient (edge) consistency. Lower scores indicate better edge preservation: values $\leq$0.05 indicate excellent gradient fidelity, 0.05–0.10 good, and above 0.10 signal degraded sharpness or more different between each signal. Mean Deviation Similarity Index (MDSI) combines color, luminance and gradient information into a single deviation measure. MDSI $\in [0, 1]$, where 0.00 is perfect. Values $\leq$0.05 denote excellent overall similarity, 0.05–0.10 good, and >0.10 poor similarity or total difference.

Overlapping estimation

In 4D-STEM data extraction, the issue of overlapping arises when diffraction patterns from adjacent sample regions

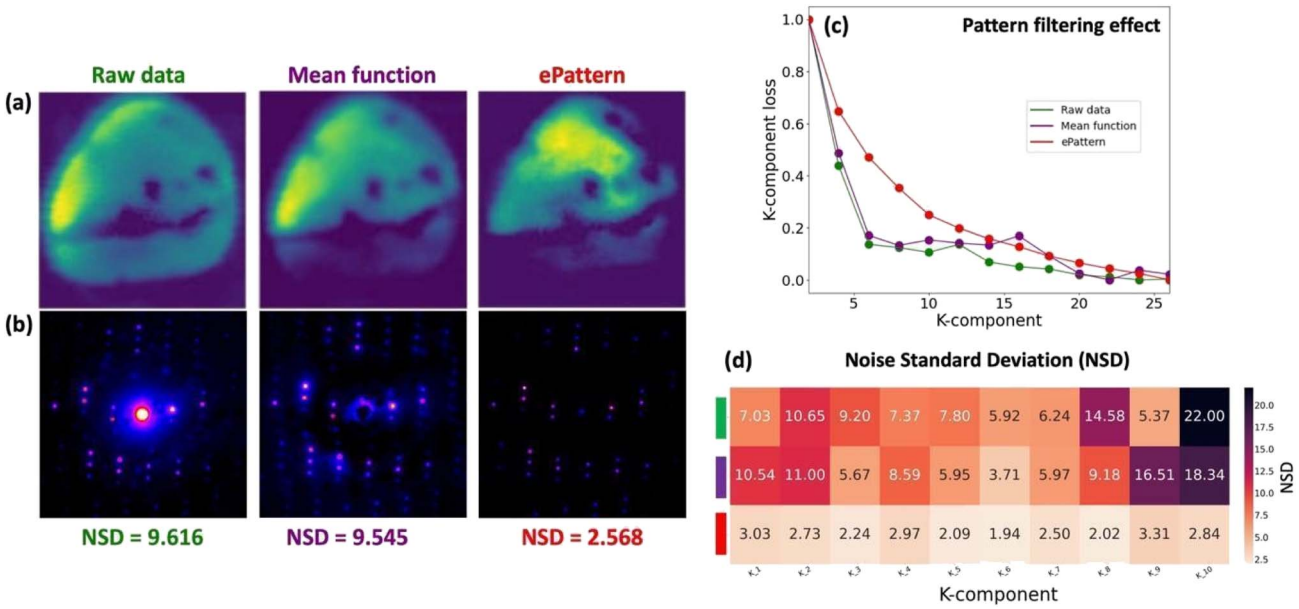

Fig. 2 Influence of dataset filtering on NMF robustness. (a and b) Orientation maps (top) and representative diffraction patterns (bottom) for raw data, mean function filtered data, and ePattern-processed data. NSD values quantify noise reduction performance. (c) *K*-Component loss curves for each dataset, illustrating convergence stability with the growth of components (clustering number), the *K*-component loss is defined as an evaluated method to measure the similarity ($V - V'$) between the dataset that needs to be dealt with NMF (V) and its corresponding NMF result ($V' = W \times H$). (d) Heatmap of NSD values across clusters (rows) and datasets (columns), with darker hues indicating higher noise.

interfere, leading to the superposition of signals during data acquisition.[24,39] Diffraction patterns generated at contiguous positions inherently contain contributions from neighboring areas of the sample.[26] This overlap can complicate data interpretation, particularly when analyzing subtle structural features. The primary challenge posed by overlapping in 4D-STEM lies in accurately extracting information related to the sample's local crystallography and structural properties.[39] The reliability of the reconstructed patterns heavily depends on the precision of the analytical techniques employed. Overlapping signals can introduce artifacts or inaccuracies, potentially compromising the fidelity of the final results.

4D-STEM data are first transformed into a 2D array and subsequently processed using NMF to obtain the W and H matrices. The H matrix ($k, M \times N$) represents the contribution of each basis vector from the W matrix to reconstructing the original matrix V.[40] Each row of H corresponds to the weights for specific data points in V, reflecting the extent to which each basis vector contributes to the reconstruction. In this context, H encapsulates how the features represented by W are combined to describe the original data. Here, k denotes the number of clusters within the dataset, with each column indicating the weight (or probability) of a data point (diffraction pattern) belonging to a given cluster. Higher weights signify a greater likelihood of association with a specific cluster.

NMF exhibits a strong sensitivity to pertinent features during matrix factorization, effectively capturing overlapping structures within the dataset. In this context, overlaps are represented as secondary weights, while the original clusters correspond to primary weights. Utilizing the H matrix, which encapsulates all weight contributions, we extract the first and second weights and define a threshold to differentiate them. This threshold facilitates the visualization of overlapping regions within the dataset, as shown in Fig. 5.

Comparison of raw and clustered data

Clustering analysis using NMF proves highly effective for 4D-STEM data, as it reveals latent features associated with distinct structural characteristics.[25,41] Beyond identifying hidden patterns, NMF generates interpretable clustering results that correlate with specific crystallographic behaviors, enhancing the understanding of complex material structures.[27] Integrating clustering outputs with the original diffraction patterns not only provides validation but also validates the making-decision step for *k*-component determination. Indeed, aligning the original diffraction data with NMF-derived clusters strengthens the validation process, enabling a more robust assessment of cluster coherence relative to material properties and structural variations. This dual representation facilitates intuitive visualization, bridging the gap between mathematical models and structural information. Moreover, correlating NMF clusters with specific regions within the raw data allows for precise localization of structural features, thereby enriching the interpretability of clustering outcomes.

Results and discussion

Impact of dataset filtering on NMF robustness

Applying appropriate dataset filtering techniques can significantly reduce noise, thereby enhancing the robustness of NMF.[7,9] By eliminating noisy data points, filtering ensures that NMF focuses on extracting underlying patterns rather than capturing

noise, resulting in more reliable factorization outcomes.[8] The experimental impact of dataset filtering on NMF is illustrated in Fig. 2. Fig. 2a demonstrates the substantial influence of filtering methods on the clustering quality of NMF results.

Three different filtering approaches are compared: (1) raw data: the dataset as acquired from 4D-STEM without any processing. (2) Mean filtering: this method processes the raw data by normalizing the sum of neighboring images using a 3×3 kernel sliding across the scan.[42] This averaging technique produces a scan of unchanged size, where each Diffraction Pattern (DP) image is the average of neighboring images.[42] (3) ePattern algorithm: proposed in our team, this novel algorithm focuses on dimensionality reduction and reconstruction of DP.[42] It employs a neural network-like structure consisting of an encoder, which extracts the most relevant features into a latent space, and a decoder, which reconstructs the diffraction patterns from the latent space representation.[42] These filtering methods highlight the importance of preprocessing in enhancing the quality and reliability of NMF results, particularly in the context of 4D-STEM data analysis.

The two proposed methods aim to enhance data quality by eliminating noisy or irrelevant data points from the dataset. Fig. 2a illustrates the Noise Standard Deviation (NSD) values corresponding to each representative pattern extracted using NMF. Among the evaluated datasets, the ePattern dataset demonstrates the lowest NSD value (2.568), indicating that low-variance features have been effectively removed. Fig. 2c further compares the NSD values across various K-component clusters ($\text{Cluster}_1 \sim \text{Cluster}_k$) for different methods. The intensity of the heatmap corresponds to the magnitude of NSD, with ePattern consistently showing the lowest values (depicted by red and lightest colors). This reduction in noise enables NMF to achieve more efficient factorization and enhances the interpretability of the resulting components.

Fig. 2b visualizes the impact of dataset filtering on convergence and computational efficiency during NMF processing. A comparison between raw data and preprocessed datasets (mean function and ePattern) highlights the advantages of the latter. The ideal factorization result ($V' = W \times H$) is closer to the original matrix (V), with minimal deviation. Notably, the loss curve for the ePattern dataset exhibits a smooth and consistent downward trend, unlike the raw data and mean function, which show numerous outliers. At the critical point of the steepest gradient change ($k = 10$), the ePattern curve demonstrates minimal fluctuation, underscoring its stability and robustness in noise handling. This improved convergence behavior facilitates more reliable and accurate NMF performance.

Moreover, by removing irrelevant features, the ePattern dataset enables NMF to produce more interpretable factors. When applied to the ePattern dataset, the resulting components represent distinct and meaningful patterns that are easier to interpret and analyze (Fig. 5). In addition to superior noise reduction, the ePattern dataset enhances the stability of NMF, reduces the risk of overfitting, and prevents the model from capturing artificial patterns originating from noise.[45]

Influence of IQA on determining the optimal K component in NMF analysis

Determining the optimal number of components k in NMF requires balancing the trade-off between reconstruction quality and model complexity.[19,21] This involves evaluating image quality assessment (IQA) metrics and reconstruction loss across different values of k, with the objective of identifying the optimal value that provides a faithful approximation of the original data while avoiding unnecessary complexity.[46] The overarching aim is to find a value of k that effectively captures the underlying structure of the data while maintaining computational efficiency.[41]

When applied to image clustering, the quality of the reconstructed images and the accuracy of the decomposition are pivotal in determining the optimal k.[22,23] IQA metrics, are utilized to evaluate the fidelity and differences in the reconstructed images. The reconstructed data V' is expressed as $V' = W \times H$, where each clustering operation corresponds to ($\text{Clustering}_1 = W_1 \times H_1 \ldots \text{Clustering}_k = W_k \times H_k$).

Fig. 3a demonstrates that increasing k generally reduces reconstruction loss, as a larger number of components can theoretically capture more details of the original data. However, this also introduces the risk of overfitting, where the model begins to capture noise along with the signal. Higher values of k tend to improve IQA metrics, such as SSIM, up to a threshold, after which additional components may not enhance quality and might even degrade it due to overfitting.

The range of interest identified in Fig. 3a suggests that k values between approximately 6 and 14 (centered around $k = 10$) achieve an optimal balance between underfitting and overfitting. Within this range, reconstruction loss decreases significantly while avoiding overfitting. This range also reflects a trade-off between capturing essential features and minimizing the incorporation of noise.

Fig. 3b–e analyze k based on four IQA algorithms. For instance, Fig. 3e examines PSNR, a metric used to measure the fidelity of reconstructed images by comparing them with the original. Higher PSNR values indicate reduced distortion and noise, signifying that the NMF components have effectively captured the essential features of the original data.[34] In image compression and reconstruction contexts, PSNR values above 40 are considered excellent, whereas values below 20 are deemed unacceptable.[34] For NMF, PSNR values higher than 40 indicate that the reconstructed images retain a high degree of similarity to the original data, which is crucial for determining the optimal k.

The results suggest that $k = 10$ (or slightly below this value) achieves an equilibrium between preserving essential features and avoiding noise overfitting. While PSNR provides a global perspective on the fidelity of image reconstruction, other metrics like MDSI, GMSD, and SSIM complement the analysis by focusing on different aspects of image quality.

Fig. 3c presents the results of MDSI, which evaluates global differences between images, including intensity and spatial information.[47] MDSI values range from 0 to 1, with higher values indicating greater similarity.[47] For NMF clustering, the

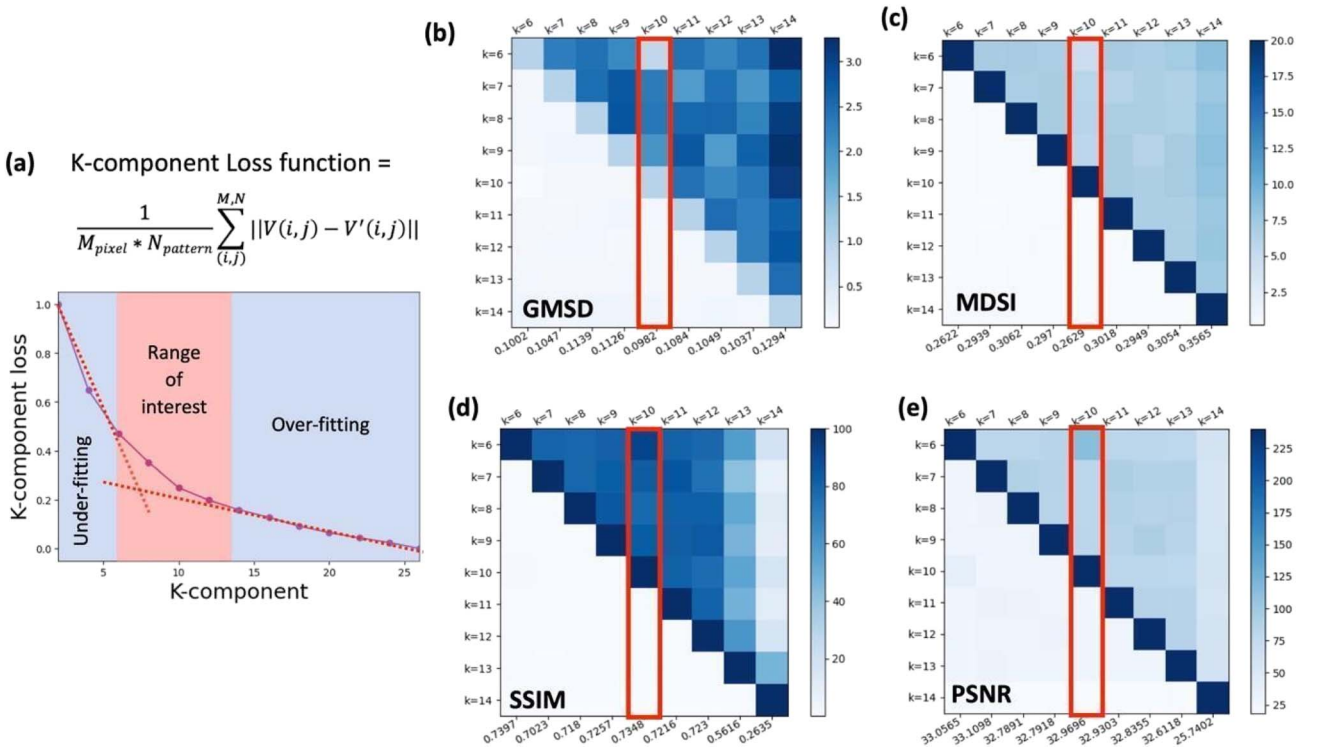

Fig. 3 Mean absolute error (MAE) between the original dataset (V) and NMF-reconstructed dataset (V') as a function of component count (k). The curve identifies the optimal k (here, $k = 8$), where error reduction balances reconstruction fidelity and model simplicity. (a) The figure demonstrates the tendency of decline according to the increase of component (k), the determination of the range of interest where the ideal component presumably existed in this. (b–d) The diffraction pattern selected by the most intense value between the $k - 4$ and $k + 4$ clustering is used for calculating IQA loss for decision-making two. The three matrices represent the result of the evaluation of each current diffraction pattern with the others from the decision-making one through three different IQA algorithms in SSIM, GMSD and MDSI. The objective of all metrics is to measure the similarity between two images, which aim is to determine the number of most different orientation pattern forms, namely the ideal component. (e) PSNR is objective to assess the degree of alteration/degradation between the diffraction pattern in the current clustering and the one in the other clustering. The result of PSNR is higher, signifying the more pertinent contribution among the totality of components.

goal is to maximize the distinctiveness of clusters, ensuring that diffraction patterns within clusters are noticeably distinct. At $k = 8$, MDSI (value = 0.3062) captures global features effectively while minimizing distortion.

In contrast, Fig. 3b and d focus on GMSD and SSIM, which measure localized and structural differences. GMSD quantifies deviations in gradient magnitudes between reference and reconstructed images, making it suitable for capturing changes in image structure caused by distortions.[48,49] Lower GMSD values indicate greater similarity, while higher values highlight increased dissimilarity.[50] At $k = 8$, GMSD achieves an ideal value of 0.1139, signifying effective structural fidelity.

Concurrently, the SSIM provides a robust evaluation of the similarity between two images by assessing their structural information.[46] Notably, SSIM is highly sensitive to subtle structural differences, making it an effective tool for detecting slight variations between images. The SSIM index ranges from −1 to 1, where a value of 1 represents perfect structural similarity, and −1 indicates complete dissimilarity.[46,51]

In the context of clustering optimization, the analysis aims to minimize redundancy among clustering points on a global scale, with the objective of maximizing the sum of distinctly different clustering points.[52,53] As illustrated in Fig. 3b and d, based on the ultimate values for GMSD = 0.1139 and SSIM = 0.718, the analysis indicates that $k = 8$ represents an optimal choice for k, as further validated in Fig. 4.

Advanced analysis and interpretation of orientation mapping in 4D-STEM *via* NMF

In the domain of 4D-STEM, NMF serves as a powerful computational tool for decomposing diffraction pattern data into distinct structural and orientation components.[27] In Fig. 4 (and Fig. SI_7), 4D-STEM analysis has been performed on LMNO cathode materials of a Li-ion battery, revealing a distinct agglomeration of crystals with noticeable overlapping between individual crystallites. This crystal configuration is inherently challenging to analyze due to the projection effects intrinsic to the TEM technique. This study emphasizes the determination of the optimal number of components (k) necessary to effectively capture and map crystallographic orientations within the dataset. By applying NMF with an optimized $k = 8$, the method successfully delineates and clusters distinct crystallographic orientations and phases.

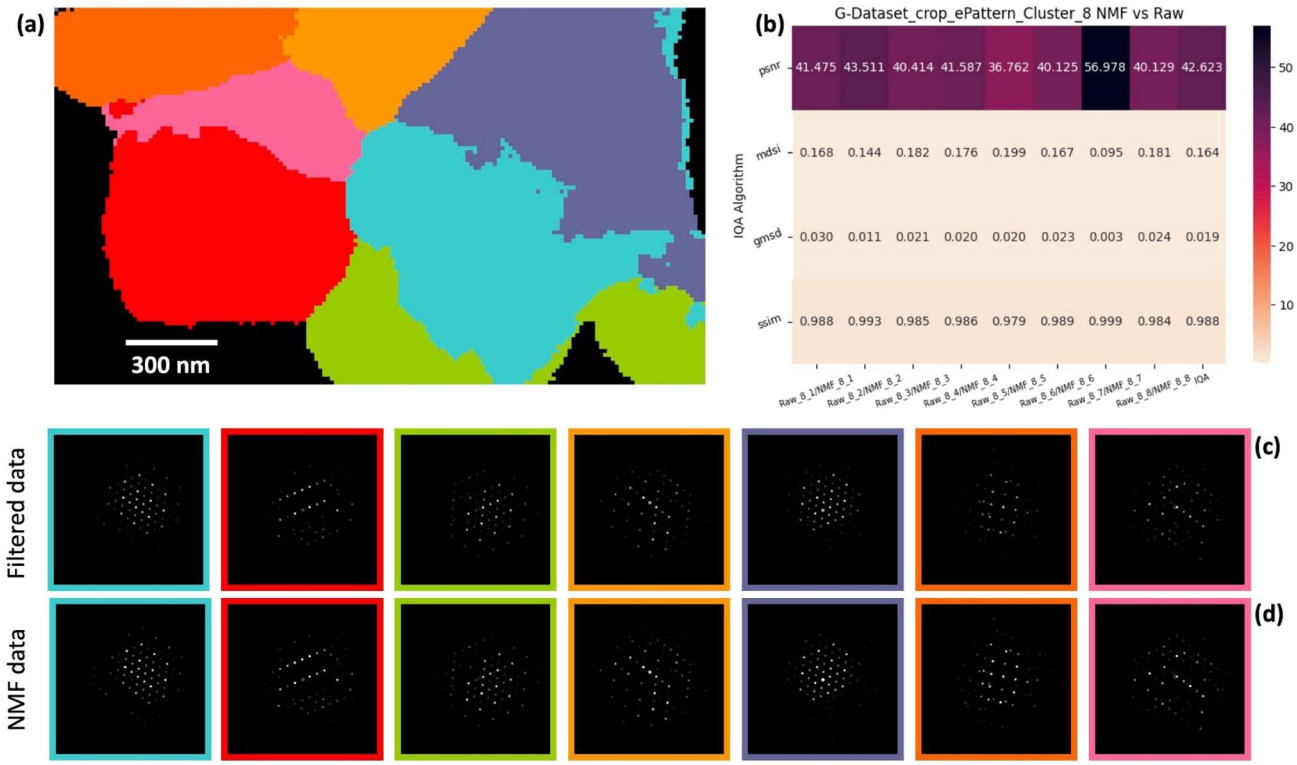

Fig. 4 Orientation mapping results and diffraction data analysis using Non-negative Matrix Factorization (NMF) applied to 4D-STEM datasets of LMNO cathode materials of Li-ion battery. (a) Reconstructed orientation map over a 2 μm region, highlighting spatial variations in grain orientations. (b) Quantitative comparison of orientation contributions across the dataset using various NMF algorithms, including PSNR, MDSI, GMSD, SSIM values. (c) Selected diffraction patterns from filtered data (ePattern), corresponding to identified orientation clusters. (d) Processed diffraction patterns after NMF decomposition, showing enhanced clarity for each orientation cluster.

The integrity of the 4D-STEM dataset, characterized by well-defined and distinct diffraction patterns, is paramount for achieving accurate component separation. Fig. 4 demonstrates the robustness of NMF, validated by quantitative image quality metrics, in extracting and mapping structural features in complex materials such as cathode materials, where lattice parameters of different phases can be very close to each other. The choice of $k = 8$ was guided by a systematic evaluation of the trade-off between capturing essential structural details and mitigating overfitting. The PSNR values are consistently high (mostly >40 dB), demonstrating that NMF reconstructions retain fine diffraction features with minimal distortion. The coherently high values are ranging from ~37 dB to ~57 dB, with most above 40 dB. Such high PSNR values indicate that the NMF reconstructions closely approximate the raw patterns while preserving the high-frequency details that are crucial for identifying weak Bragg reflections. Similarly, the SSIM scores are very close to 1 (0.979–0.993), confirming excellent structural similarity between NMF and raw patterns, which also demonstrates that the structural information in the diffraction patterns, particularly the form and relative intensity distribution of Bragg disks, is well preserved in the NMF outputs, despite the denoising process. Meanwhile, the MDSI and GMSD values remain low across all clusters, indicating negligible perceptual differences. For the MDSI evaluation, the low values (0.095–0.199) further support the conclusion that perceptual differences between raw and reconstructed patterns are minimal. MDSI is sensitive to contrast and luminance changes, and the low deviations here suggest that NMF maintains intensity relationships in the diffraction patterns. In parallel, in terms of GMSD, with values consistently below 0.03, GMSD confirms that the local gradient structures (edges and sharp intensity transitions in diffraction spots) are highly consistent between raw and reconstructed data. This metric is particularly relevant for diffraction analysis, where preserving the sharpness of Bragg disks is essential for accurate reciprocal space mapping.

The visualization in Fig. 4 encapsulates the outcome of NMF applied to the dataset, where diffraction patterns from various sample regions are color-coded to represent distinct components or orientations. Each region is associated with the most representative diffraction pattern derived from the clustering process, as shown in the bottom row of the figure (labeled 1 through 8). This mapping confirms that NMF differentiates regions based on their structural similarity. The distinct colors and their corresponding diffraction patterns further validate that the selected $k = 8$ captures the essential crystallographic orientations and phases in the sample.

Moreover, Fig. 4 underscores the critical role of dataset quality in enabling accurate component identification. High-quality diffraction patterns, characterized by sharp and well-defined features, enhance the ability of NMF to discern subtle

variations in orientations with local disorientations. The sensitivity of the model to structural and orientation features at $k = 8$ ensures a precise balance between capturing intricate details and minimizing noise. Consequently, this approach facilitates meaningful and reliable orientation mapping, emphasizing the synergy between advanced computational techniques and high-quality experimental data.

Fig. 4 illustrates results from NMF applied to the filtered dataset (*via* ePattern), while Fig. SI_7 shows unprocessed raw data. Comparison with the raw dataset reveals the importance of preprocessing: unprocessed patterns suffer from noise, obscuring weak reflections and complicating segmentation. Filtering enhances diffraction spot visibility, reduces background, and improves both interpretability and clustering accuracy.

NMF on the filtered data yields clearer, more consistent reconstructions than on raw inputs. IQA metrics confirm this: PSNR values (~37–57 dB, mostly >40 dB) show reconstructions approximate raw patterns while suppressing noise, SSIM scores (0.979–0.993) indicate strong preservation of Bragg disk features and low MDSI (0.095–0.199) and GMSD (<0.03) values show minimal perceptual or gradient differences.

Filtered clustering maps display sharper domain boundaries and better phase separation than noisy raw maps, directly improving structural insight. Overall, dataset reduction, through filtering and NMF, balances denoising with structural fidelity, ensuring both visual clarity and quantitative reliability for tasks such as strain mapping, orientation classification, and phase identification. It is thus a prerequisite for extracting robust physical insights from 4D-STEM *via* unsupervised clustering.

Identification of crystallite overlapping region

Fig. 5 demonstrates the application of NMF for the decomposition of 4D-STEM data into spatial weight matrices corresponding to individual clusters. This technique facilitates the identification and visualization of regions exhibiting significant cluster overlaps by analyzing the ratio between the second highest and maximum weights at each spatial position. Such an approach provides critical insights into the spatial distribution and interaction of clusters within the dataset.

In the context of NMF, where $V = W \times H$ encodes the weight information for each cluster. Each element $H(i,j)$ represents the probability of a specific pixel belonging to a given cluster. By reshaping H into k individual weight matrices (H_1, H_2, …, H_k), each matrix corresponds to a unique cluster and captures its spatial distribution as a 2D representation with dimensions (x, y).

To evaluate cluster overlap, the method systematically compares the maximum weight and the second-highest weight at each pixel location across all clusters. A ratio is computed as second weight/first weight, with thresholding parameters ranging from 75% to 95% to delineate regions where the second-highest weight contributes significantly. This enables the detection of areas where clusters are not well-separated, highlighting potential overlaps.

Fig. 5a illustrates the structure of H, represented as a matrix with dimensions (k, $x \times y$), where k is the number of clusters and $x \times y$ represents the flattened spatial dimensions of the dataset. For each spatial position (x, y), a corresponding weight vector in H indicates the likelihood of that position belonging to each cluster. For instance, if $H(1, 1) = 0.5$, it signifies that the

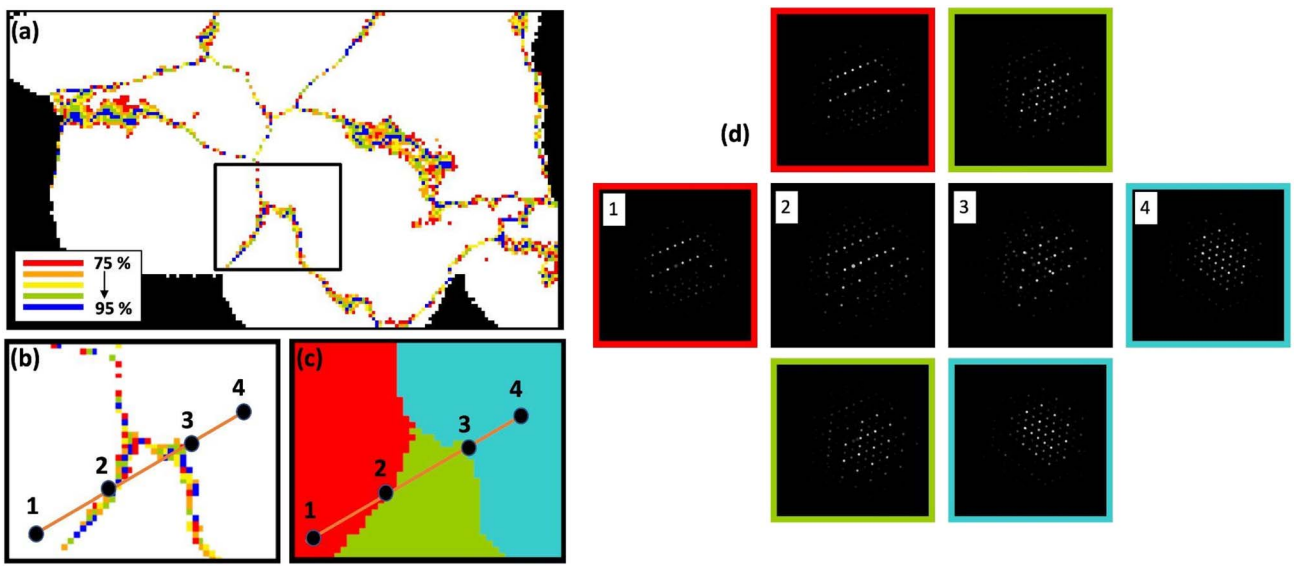

Fig. 5 (a) Spatial mapping of overlapping regions identified by applying second-weight thresholds of 75%, 80%, 85%, 90%, and 95%. Color-coded overlays highlight regions of significant overlap, with increasing thresholds represented by red (75%), orange (80%), yellow (85%), and blue (95%). (b) Magnified view of the boxed region in (a), showing the beam scanning trajectory from point 1 to point 4 across an overlapping region. (c) Corresponding thresholded cluster maps, where distinct colors indicate contributions from different clusters. These maps illustrate how the second-highest cluster weight contributes to spatial complexity at cluster boundaries. (d) Diffraction patterns at points 1–4, extracted along the scanning trajectory. The results demonstrate that overlapping regions can be decomposed into two or more distinct diffraction patterns, especially at the cluster boundary.

first diffraction pattern has a 50% likelihood of belonging to the first cluster, while $H(k, 1)$ reflects the probability of the same diffraction pattern belonging to the k-th cluster.

Following the reshaping of H into individual cluster weight matrices, Fig. 5b displays these matrices ($Cluster_1$, $Cluster_2$, …, $Cluster_k$), each showing weights specific to a single cluster. Threshold values of 75%, 80%, 85%, 90%, and 95% are applied to identify regions of significant overlap. The corresponding spatial regions are then visualized using color-coded overlays to represent varying degrees of overlap.

In Fig. 5c, the resulting map highlights regions of cluster overlap based on the second-weight thresholding. Different colors denote the degree of overlap, with red (75%), orange (80%), yellow (85%), and blue (95%) representing increasing thresholds. This visualization clearly delineates areas where the second-highest cluster weight plays a significant role, providing critical insights into the spatial complexity and potential interactions between clusters within the dataset. This section highlights the application of NMF to decompose 4D-STEM data, resolve cluster overlaps using second-to-maximum weight ratios, and visualize spatial interactions through color-coded maps.

As shown in Fig. SI_8, the results in Fig. 5 highlight that preprocessing, through denoising pre-treatment, is an indispensable step in the workflow. It suppresses noise, preserves meaningful secondary contributions, and enables the decomposition of overlapping diffraction patterns into distinct components. This treatment transforms ambiguous boundary regions into valuable sources of information, thereby allowing a more robust and physically meaningful analysis of structural complexity.

Comparing NMF results with the raw dataset

To assign each input dataset element to its corresponding diffraction cluster based on its index *via* NMF algorithm, the computed matrix H provides critical information about cluster membership. Specifically, for each column in H, $H_{(k,j)} > H_{ij}$ for all $i \neq k$, this indicates that the input data point V_j belongs to the k-th cluster.[40] Furthermore, the computed matrix W represents the cluster centroids, where the k-th column corresponds to the centroid of the k-th cluster.[40] Consequently, each diffraction pattern in the original dataset can be uniquely associated with a specific cluster index.

For instance, if $H_{(1,1)} > H_{(i,1)}$ ($i = 2, 3, 4, \ldots, k$), this indicates that the first diffraction pattern, located at position (1, 1) in the original dataset, belongs to the first cluster. Using this approach, all diffraction patterns associated with a given cluster can be identified and subsequently organized into a 3D array, where each layer corresponds to an individual diffraction image. For example, if there are N diffraction patterns of dimensions 512×512 in the first cluster, the resulting array will have dimensions $512 \times 512 \times N$.

To analyze these diffraction patterns further, the mean pixel intensity can be computed at each position (i, j) across all images in the cluster. This involves averaging the pixel values at position (i,j) across all NN diffraction patterns. Mathematically, the mean intensity at position (i, j) is given by:

$$\text{Average}_{(i,j)} = \frac{1}{N}\sum_{r=1}^{N} I_r(i,j)$$

where $I_r(i,j)$ is the pixel value at position (i,j) in the r-th image.

Similarly, in the results of NMF, the element with the maximum weight in each column $H(k, j)$ is identified. This maximum weight is then used to scale its corresponding column $W(i, k)$. The resulting products are employed to reconstruct the diffraction pattern for the current clustering k. This process is repeated for all diffraction patterns within the current clustering, yielding a new diffraction pattern that encapsulates the characteristic information of that clustering.

Returning to the original dataset enables a comparative analysis between the initial orientations and the NMF results (see in SI). This comparison not only validates the proposed method but also reinforces its effectiveness (Fig. 4). Furthermore, it contributes to a deeper understanding of material characterization within the framework of clustering analysis.

Conclusion

In this paper, we have demonstrated a robust and systematic approach to determining the optimal number of components (k) in non-negative matrix factorization (NMF) for the analysis of 4D-STEM datasets, emphasizing the critical interplay between data quality, clustering outcomes, and computational efficiency. Through the application of various image quality assessment (IQA) metrics, including PSNR, MDSI, GMSD, and SSIM, our analysis highlights how the trade-off between reconstruction fidelity and model complexity can be effectively managed to achieve an optimal k value, with $k = 8$ striking the right balance between capturing essential data features and avoiding overfitting.

The integration of unsupervised multi-clustering strategies is pivotal in this context, as it facilitates a nuanced understanding of overlapping cluster structures inherent in 4D-STEM datasets. By analyzing spatial weight matrices and applying threshold-based visualization techniques, this study identified regions with significant overlap, thus enabling the identification of interaction zones and structural patterns within the data. These insights provide a more granular perspective of cluster distributions and inter-cluster relationships, which are crucial for refining decision-making processes in NMF-based analysis pipelines.

Moreover, this study underscores the importance of data preprocessing in enhancing the robustness and interpretability of unsupervised clustering results. Three preprocessing methods, raw data, mean function, and ePattern, were evaluated, with the ePattern method yielding the most consistent and reliable outcomes by significantly reducing noise (lower NSD values) and removing low-variance features. This demonstrates that high-quality datasets not only improve the stability of NMF results but also enable more effective multi-clustering strategies by focusing on meaningful data patterns.

Decision-making strategies in this study were further strengthened by employing IQA metrics as quantitative tools to guide the determination of k. The metrics reveal that while

higher k values initially improve reconstruction accuracy, there is a threshold beyond which additional components contribute negligible quality improvements and risk overfitting. This informed decision-making approach ensures that NMF-derived results remain both computationally efficient and scientifically interpretable.

In conclusion, our study highlights a comprehensive framework that combines dataset preprocessing, unsupervised multi-clustering, and decision-making strategies to optimize NMF-based analysis of 4D-STEM datasets. By addressing overlapping cluster structures and leveraging data quality enhancements, this methodology not only improves the robustness and reliability of factorization results but also provides actionable insights into complex structural properties of cathode crystals in the 4D-STEM data. These findings establish a foundational approach for future research leveraging NMF in complex, multi-dimensional datasets and reinforce the significance of systematic preprocessing and decision-making frameworks in achieving reliable and interpretable outcomes.

Conflicts of interest

The authors declare no competing financial or non-financial interests.

Data availability

The 4D-STEM datasets used in this contribution is available for free download at **https://doi.org/10.5281/zenodo.15492699**.

Code availability: The ePattern_Clustering is available for free download at **https://doi.org/10.5281/zenodo.17214464**.

Supplementary information (SI): provides detailed descriptions of the Non-Negative Matrix Factorization (NMF) algorithm, the dataset processing workflow from 4D to 2D, and the calculation methods for the Noise Standard Deviation (NSD), Peak Signal-to-Noise Ratio (PSNR), Mean Deviation Similarity Index (MDSI), Gradient Magnitude Similarity Deviation (GMSD), and Structural Similarity Index (SSIM). It also includes the global scheme of the NMF-clustering analysis and the implementation details of the ePattern algorithm. See DOI: **https://doi.org/10.1039/d5dd00071h**.

Acknowledgements

The research presented in this paper has received support from multiple sources. Specifically, funding has been provided by the French Research Agency (ANR) as part of the DestiNation_operando project (ANR-19-CE42-0014) and by the company NanoMegas (Belgium). Additionally, the UPJV and RS2E electron microscopy platforms were utilized for this research. The authors express their gratitude to Fayçal Adrar (LRCS/RS2E) for his invaluable assistance in the testing and comparative analysis of various models in 4D-STEM data processing.

References

1 (*a*) F. Uesugi, S. Koshiya, J. Kikkawa, T. Nagai, K. Mitsuishi and K. Kimoto, Non-negative matrix factorization for mining big data obtained using four-dimensional scanning transmission electron microscopy, *Ultramicroscopy*, 2021, **221**, 113168, DOI: **10.1016/j.ultramic.2020.113168**; (*b*) K. C. Bustillo, S. E. Zeltmann, M. Chen, J. Donohue, J. Ciston, C. Ophus and A. M. Minor, 4D-STEM of beam-sensitive materials, *Acc. Chem. Res.*, 2021, **54**(11), 2543–2551.

2 (*a*) E. F. Rauch, J. Portillo, S. Nicolopoulos, D. Bultreys, S. Rouvimov and P. Moeck, Automated nanocrystal orientation and phase mapping in the transmission electron microscope on the basis of precession electron diffraction, *Z. Kristallogr. – Cryst. Mater.*, 2010, **225**(2–3), 103–109, DOI: **10.1524/zkri.2010.1205**; (*b*) E. F. Rauch and M. J. M. C. Véron, Automated crystal orientation and phase mapping in TEM, *Mater. Charact.*, 2014, **98**, 1–9; (*c*) B. H. Savitzky, S. E. Zeltmann, L. A. Hughes, H. G. Brown, S. Zhao, P. M. Pelz, et al.). py4DSTEM: A software package for four-dimensional scanning transmission electron microscopy data analysis, *Microsc. Microanal.*, 2021, **27**(4), 712–743; (*d*) C. Francis and P. M. Voyles, pyxem: A Scalable Mature Python Package for Analyzing 4-D STEM Data, *Microsc. Microanal.*, 2023, **29**, 685–686.

3 M. Ringnér, What is principal component analysis?, *Nat. Biotechnol.*, 2008, **26**(3), 303–304.

4 I. T. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments, *Philos. Trans. R. Soc., A*, 2016, **374**(2065), 20150202.

5 T. Kurita, Principal component analysis (PCA), *Computer Vision: A Reference Guide*, 2019, pp. 1–4.

6 S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani and A. Hooman, An overview of principal component analysis, *J. Signal Inf. Process.*, 2013, **4**(3B), 173.

7 T. M. Mitchell and P. Langley, *Machine Learning*, McGraw-Hill, New York, 1997, vol. 1, p. 9.

8 J. Hu, H. Niu, J. Carrasco, B. Lennox and F. Arvin, Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning, *IEEE Trans. Veh. Technol.*, 2020, **69**(12), 14413–14423, DOI: **10.1109/TVT.2020.3034800**.

9 T. Hastie, R. Tibshirani and J. Friedman, Unsupervised Learning, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ed. T. Hastie, R. Tibshirani and J. Friedman, Springer New York, New York, NY, 2009, pp. 485–585, DOI: **10.1007/978-0-387-84858-7_14**.

10 (*a*) G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, 2020; (*b*) S. Zhou, Z. Xu and F. Liu, Method for determining the optimal number of clusters based on agglomerative hierarchical clustering, *IEEE Trans. Neural Networks Learn. Syst.*, 2016, **28**(12), 3007–3017; (*c*) A. Kuraria, N. Jharbade and M. Soni, Centroid selection process using WCSS and elbow method for K-mean clustering algorithm in data mining, *Int. J. Sci. Res. Sci. Eng. Technol.*, 2018, **12**, 190–195; (*d*) A. M. Bagirov,

R. M. Aliguliyev and N. Sultanova, Finding compact and well-separated clusters: Clustering using silhouette coefficients, *Pattern Recognit.*, 2023, **135**, 109144; (*e*) X. Hu, Y. Tang, W. Pedrycz, J. Jiang and Y. Jiang, Knowledge-Driven Possibilistic Clustering with Automatic Cluster Elimination, *Comput., Mater. & Continua*, 2024, **80**(3), 4917–4945.

11 T. Kurita, Principal component analysis (PCA), *Computer Vision: A Reference Guide*, 2019, pp. 1–4.

12 M. Ringnér, What is principal component analysis?, *Nat. Biotechnol.*, 2008, **26**(3), 303–304.

13 E. C. Malthouse, Limitations of nonlinear PCA as performed with generic neural networks, *IEEE Trans. Neural Netw.*, 1998, **9**(1), 165–173.

14 E. N. Brockmann and W. P. Anthony, Tacit knowledge and strategic decision making, *Group Organ. Manag.*, 2002, **27**(4), 436–455.

15 E. Triantaphyllou and E. Triantaphyllou, *Multi-Criteria Decision Making Methods*, Springer, 2000.

16 T. Hastie, R. Tibshirani, J. H. Friedman and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009, vol. 2.

17 G. W. Milligan and M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, 1985, **50**(2), 159–179, DOI: **10.1007/BF02294245**.

18 J. Oyelade, et al., Data Clustering: Algorithms and Its Applications, in *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*, 2019, pp. 71–81, DOI: **10.1109/ICCSA.2019.000-1**.

19 R. Tibshirani, G. Walther and T. Hastie, Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 2001, **63**(2), 411–423.

20 (*a*) S. J. Pennycook and P. D. Nellist, *Scanning Transmission Electron Microscopy: Imaging and Analysis*, Springer Science & Business Media, 2011; (*b*) B. H. Martineau, D. N. Johnstone, A. T. van Helvoort, P. A. Midgley and A. S. Eggeman, Unsupervised machine learning applied to scanning precession electron diffraction data, *Adv. Struct. Chem. Imaging*, 2019, **5**, 1–14.

21 P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, 1987, **20**, 53–65.

22 R. Tibshirani, G. Walther and T. Hastie, Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 2001, **63**(2), 411–423.

23 D. M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci.*, 2004, **44**(1), 1–12.

24 C. Ophus, Four-dimensional scanning transmission electron microscopy (4D-STEM): From scanning nanodiffraction to ptychography and beyond, *Microsc. Microanal.*, 2019, **25**(3), 563–582.

25 (*a*) T. Friedrich, C.-P. Yu, J. Verbeeck and S. Van Aert, Phase object reconstruction for 4D-STEM using deep learning, *Microsc. Microanal.*, 2023, **29**(1), 395–407; (*b*) A. Valery, E. F. Rauch, L. Clément and F. Lorut, Retrieving overlapping crystals information from TEM nano-beam electron diffraction patterns, *J. Microsc.*, 2017, **268**(2), 208–218.

26 C. Ophus, et al., Automated crystal orientation mapping in py4DSTEM using sparse correlation matching, *Microsc. Microanal.*, 2022, **28**(2), 390–403.

27 (*a*) F. I. Allen, et al., Fast Grain Mapping with Sub-Nanometer Resolution Using 4D-STEM with Grain Classification by Principal Component Analysis and Non-Negative Matrix Factorization, *Microsc. Microanal.*, 2021, **27**(4), 794–803, DOI: **10.1017/S1431927621011946**; (*b*) T. Printemps, K. Dabertrand, J. Vives and A. Valery, Application of a novel local and automatic PCA algorithm for diffraction pattern denoising in TEM-ASTAR analysis in microelectronics, *Ultramicroscopy*, 2024, **267**, 114059.

28 K. Allab, L. Labiod and M. Nadif, Simultaneous semi-NMF and PCA for clustering, in *2015 IEEE International Conference on Data Mining*, IEEE, 2015, pp. 679–684.

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

30 Y.-X. Wang and Y.-J. Zhang, Nonnegative Matrix Factorization: A Comprehensive Review, *IEEE Trans. Knowl. Data Eng.*, 2013, **25**(6), 1336–1353, DOI: **10.1109/TKDE.2012.51**.

31 P. M. Kroonenberg and J. De Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, *Psychometrika*, 1980, **45**, 69–97.

32 R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 1996, **58**(1), 267–288.

33 R. Tibshirani, The lasso method for variable selection in the Cox model, *Stat. Med.*, 1997, **16**(4), 385–395.

34 D. R. Bull and F. Zhang, Digital picture formats and representations, in *Intelligent Image and Video Compression*, ed. D. R. Bull and F. Zhang, Academic Press, Oxford, 2nd edn, 2021, ch. 4, pp. 107–142, DOI: **10.1016/B978-0-12-820353-8.00013-X**.

35 H. R. Sheikh and A. C. Bovik, Image information and visual quality, *IEEE Trans. Image Process.*, 2006, **15**(2), 430–444.

36 N. Burningham, Z. Pizlo and J. P. Allebach, Image quality metrics, *Encyclopedia of Imaging Science and Technology*, 2002, vol. 1, pp. 598–616.

37 D. Jayaraman, A. Mittal, A. K. Moorthy and A. C. Bovik, Objective quality assessment of multiply distorted images, in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, pp. 1693–1697, DOI: **10.1109/ACSSC.2012.6489321**.

38 M. Shahid, A. Rossholm, B. Lövström and H.-J. Zepernick, No-reference image and video quality assessment: a classification and review of recent approaches, *EURASIP J. Image Video Process.*, 2014, **2014**, 1–32.

39 P. M. Pelz, I. Johnson, C. Ophus, P. Ercius and M. C. Scott, Real-time interactive 4D-STEM phase-contrast imaging from electron event representation data: Less computation with the right representation, *IEEE Signal Process. Mag.*, 2021, **39**(1), 25–31.

40 C. Ding, X. He and H. D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, in *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, 2005, pp. 606–610.
41 G. Chennupati, R. Vangara, E. Skau, H. Djidjev and B. Alexandrov, Distributed non-negative matrix factorization with determination of the number of latent features, *J. Supercomput.*, 2020, **76**, 7458–7488.
42 N. Folastre, et al., Improved ACOM pattern matching in 4D-STEM through adaptive sub-pixel peak detection and image reconstruction, *Sci. Rep.*, 2024, **14**, 12385.
43 D. K. Lee, J. In and S. Lee, Standard deviation and standard error of the mean, *Korean J. Anesthesiol.*, 2015, **68**(3), 220.
44 V. V. Lukin, et al., Testing of methods for blind estimation of noise variance on large image database, *Theoretical and Practical Aspects of Digital Signal Processing in Informational-Telecommunication Systems*, 2009.
45 W. Liu, N. Zheng and Q. You, Nonnegative matrix factorization and its applications in pattern recognition, *Chin. Sci. Bull.*, 2006, **51**, 7–18.
46 Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.*, 2004, **13**(4), 600–612, DOI: **10.1109/TIP.2003.819861**.
47 H. Z. Nafchi, A. Shahkolaei, R. Hedjam and M. Cheriet, Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator, *IEEE Access*, 2016, **4**, 5579–5590.
48 W. Xue, L. Zhang, X. Mou and A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, *IEEE Trans. Image Process.*, 2013, **23**(2), 684–695.
49 A. Liu, W. Lin and M. Narwaria, Image quality assessment based on gradient similarity, *IEEE Trans. Image Process.*, 2011, **21**(4), 1500–1512.
50 G.-H. Chen, C.-L. Yang and S.-L. Xie, Gradient-based structural similarity for image quality assessment, in *2006 International Conference on Image Processing*, IEEE, 2006, pp. 2929–2932.
51 Z. Wang, A. C. Bovik and H. R. Sheikh, Structural similarity based image quality assessment, in *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2017, pp. 225–242.
52 J. Immerkaer, Fast noise variance estimation, *Comput. Vis. Image Understand.*, 1996, **64**(2), 300–302.
53 P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 1994, **5**(2), 111–126, DOI: **10.1002/env.3170050203**.