Advancing Nanobeam 4D-STEM Through Manifold Learning in Cepstrum Space

Chuqiao Shi[1], Tiancheng Yu, Aaron Bayles, Zhihua Cheng[2], Matthew R. Jones[1,2], Naomi Halas, Yimo Han[1#]

[1] Department of Materials Science and NanoEngineering, Rice University, Houston, TX

[2] Department of Chemistry, Rice University, Houston, TX, United States

These authors contributed equally: Chuqiao Shi, Tiancheng Yu.

Abstract:

Nanobeam Four-Dimensional Scanning Transmission Electron Microscopy (4D-STEM) has emerged as a powerful tool for characterizing the crystal structure of a diverse range of nanomaterials. Facing the large and unlabeled 4D-STEM datasets, unsupervised machine learning has been employed to uncover hidden patterns based on the similarities between diffractions. However, the diffraction intensity variances caused by unrelated features including the lattice mistilt and dynamic scattering effects can dominate the similarity and hinder the performance of machine learning models. In this paper, we investigated a new pre-transformation strategy for 4D datasets by combining the cepstrum transformation and manifold learning to extract relevant lattice features with physical meanings such as strain profile, lattice orientation and domain structures, thereby enhancing unsupervised machine learning performance. The effectiveness of our methods is rigorously validated using both simulated and experimental datasets encompassing various materials, including 2D ferroelectric materials and nanoparticles. These data transformation techniques hold significant promise as integral components of automated 4D-STEM processing workflows, facilitating efficient and precise characterization of nanostructured materials.

Introduction:

Characterizing the lattice structures at the nanoscale is indispensable for gaining insight into the intricate relationship between a material's structure and its properties. Scanning Transmission Electron Microscopy (STEM) is a highly effective method of characterization that finds wide application. Recent developments of pixelated and high-speed direct electron detectors, enable the acquisition of one diffraction pattern in momentum space (kx, ky) at each scanning position in real space (x, y) resulting in a dataset encompassing four dimensions, aptly termed 4D-STEM. Nanobeam 4D-STEM which uses a low convergence angle probe

to generate non-overlapping diffraction disks, can provide high-resolution structural maps of diverse material systems, including 2D materials, batteries, and nano catalysts, all within a wide-ranging micron field of view. Despite its tremendous potential, the wealth of information inherent in 4D-STEM datasets poses formidable challenges in terms of data analysis, especially when applied to novel material systems.

In recent times, the electron microscopy field has witnessed the burgeoning emergence of machine learning as a promising data processing tool due to its ability to analyze complex patterns in large datasets. While supervised deep learning has proven successful in identifying structural anomalies, such as dopant atom defects, and predicting numeric properties like lattice tilt angles or sample thickness, its efficacy is often constrained by the reliance on manually curated training datasets shaped by researchers' prior knowledge. Such datasets may inadvertently overlook unexpected features within novel materials. Conversely, unsupervised learning, which doesn't require predefined training data, has been instrumental in segmenting extensive 4D datasets and extracting novel features grounded in the similarities of diffraction intensities across diffraction patterns. Nevertheless, these diffraction intensities are susceptible to distortions arising from unrelated factors, such as dynamical diffraction and lattice mistilts, which can significantly overshadow genuine similarities and impede the effectiveness of unsupervised learning. To mitigate these challenges and improve the unsupervised learning performance, there arises a need for feature extraction methods aimed at enhancing the extraction of genuine lattice features while reducing the influence of unrelated variables.

In this paper, we introduce a two-step data pre-transformation strategy to mitigate these challenges. First we use the Exit-Wave Power-Cepstrum (EWPC) method to transform the diffraction into the cepstrum space and then map the data into a low-dimension manifold space through the Uniform Manifold Approximation and Projection (UMAP) method. Our manifold learning approach in the cepstrum space extracts relevant lattice features with physical meanings, thereby enhancing the following clustering performance. We test our method in the simulated data and also apply it on various experimental datasets, including 2D ferroelectric SnSe, Al-AlO nanoparticles and AuPd core shell nanocubes. The unexpected but significant fine structures are uncovered through clustering on the transformed 4D-STEM datasets.

Results:

The overall workflow of manifold learning in cepstrum space is delineated in Figure 1. After the initial data acquisition step, a two-fold data transformation process is employed, involving both EWPC and UMAP methods. This dual transformation is conducted to extract lattice features from the acquired data and subsequently embed these features into a three-dimensional (3D) manifold space. Following this transformation, hierarchical clustering is applied within the manifold space, and the resulting clustering labels are then mapped back to facilitate visual representation.

The initial stage of the data transformation process involves the application of EWPC (Embedding Weighted Principal Component). Cepstral analysis, initially introduced in the field of electron diffraction, serves as a robust method for strain mapping and is also instrumental in the analysis of lattice distortion and amorphous materials. Figure SI illustrates the EWPC method, employing a NBED nano cube with a [100] zone axis as an illustrative example. The first step involves the calculation of the logarithm scale of the diffraction pattern, serving to mitigate intensity variance and enhance the visibility of high-order diffraction spots with low intensity in the linear scale. Subsequently, a Gaussian mask is superimposed on the logarithmic scale diffraction pattern to mitigate edge effects. Finally, the Fast Fourier Transform is computed for the masked pattern, yielding the cepstrum image.

The primary objective of the EWPC method is to segregate distinct spatial frequencies within the diffraction pattern. The lower-frequency components of the diffraction pattern are situated at the center of the cepstrum, primarily arising from unrelated features such as multi-scattering and lattice misalignments. Conversely, the higher-frequency components of the diffraction pattern are concentrated in the outer cepstrum region, primarily originating from periodic diffraction spots housing lattice-specific information. Consequently, a ring mask is applied to the cepstrum to facilitate the segregation of different frequency information, enabling the selection of high-frequency features for subsequent analysis.

The selected features within the cepstrum images are subsequently embedded into a lower three-dimensional (3D) manifold space utilizing the UMAP method. UMAP represents a contemporary dimension reduction technique, which has found applications in the visualization of biological single-cell images and has also been introduced in the electron microscopy field for the analysis of convergence beam electron diffraction (CBED) patterns. UMAP operates by computing similarity scores among data points and aims to preserve these similarities within the lower-dimensional manifold space. Consequently, similar data points tend to cluster together, rendering UMAP highly suitable for subsequent clustering procedures.

In our study, we conducted a comparative analysis of UMAP against two commonly employed dimension reduction methodologies, namely Principal Component Analysis (PCA) and Variational Auto-Encoder

(VAE) (refer to Figure SI). PCA is a linear dimension reduction technique known for its simplicity and computational efficiency. However, PCA exhibits limitations in capturing finer-grained features within sub-clusters, primarily excelling in identifying major differences. Conversely, VAE leverages an encoder-decoder neural network architecture, facilitating the extraction of latent features invariant to certain parameters, such as rotation. However, VAE is trained to minimize a loss function, enforcing the latent space to be smooth and continuous, which can be detrimental to subsequent clustering processes. Therefore, we have opted for UMAP as it effectively transforms high-dimensional features into a lower-dimensional manifold space, striking a balance between dimensionality reduction and feature preservation in support of our clustering objectives.

Following the pre-transformation steps, hierarchical clustering is applied to the manifold space data. While the K-means method is a widely adopted clustering technique due to its computational efficiency and ability to yield reasonably satisfactory results in many cases, its applicability is limited when the underlying manifold data distribution is not characterized by clean, spherical clusters. In such scenarios, clustering methods that rely on Euclidean distance, including K-means, may struggle to effectively partition the data. As an alternative, HDBSCAN distinguishes data points based on their distribution density, thus obviating the need for predefined cluster centers or manually specified cluster counts. It proves especially beneficial when dealing with complex, non-spherical data distributions. However, it is worth noting that the HDBSCAN method is more intricate, as its clustering results are notably contingent on input parameters, which may necessitate fine-tuning and could vary across different datasets. Consequently, in the context of this study, we applied both K-Means and HDBSCAN methods within the hierarchical clustering workflow, and discussed how to choose the methods practically. Subsequent to the clustering process, the labels derived from this procedure are mapped back onto the original data space, thereby facilitating a more robust segmentation of the extensive 4D-STEM dataset. This segmentation approach, particularly when investigating disparities among distinct clusters within the cepstrum domains, aims to provide deeper insights into the underlying material structures.

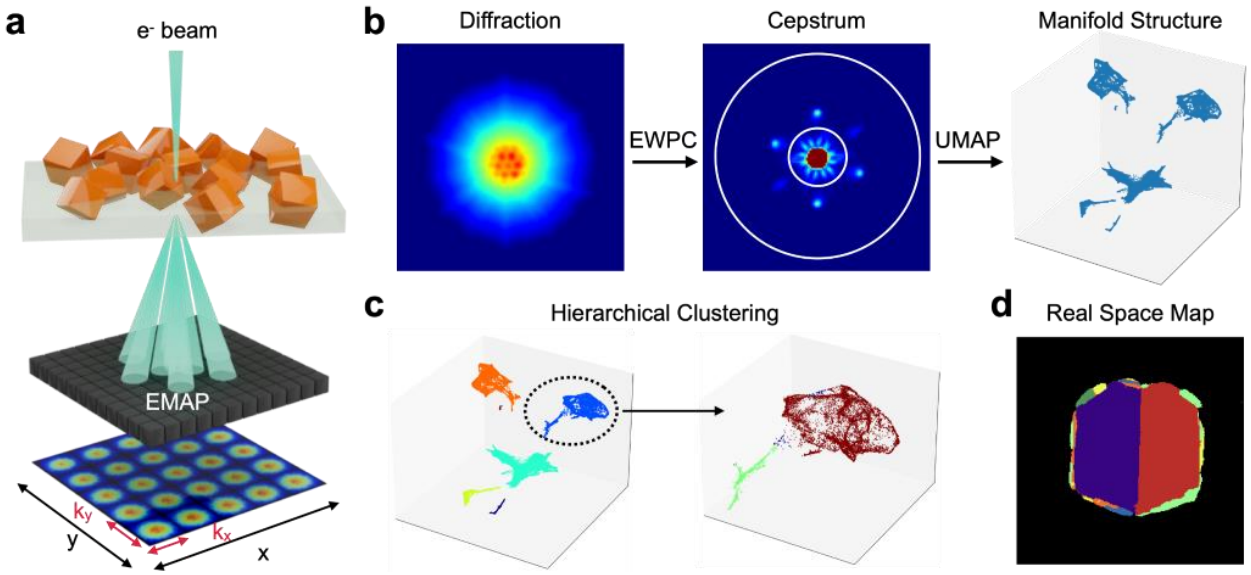

Figure 1| Overview of Manifold Learning in Cepstrum Space for 4D-STEM Datasets. (a) Illustration of the 4D-STEM data acquisition process, with EMPAD recording one diffraction pattern at each scanning position. **(b)** Data pre-transformation: The initial transformation involves converting diffraction patterns to Cepstrum space using the EWPC method. Subsequently, the high-frequency region in the Cepstrum image, denoted by the white ring, is selected, and these features are further embedded into a 3D manifold space using the UMAP method. **(c)** Hierarchical clustering in the manifold space: The complete dataset undergoes partitioning into multiple clusters, with subsequent rounds of clustering for sub-regions within the sample. **(d)** Real-space label map: Following hierarchical clustering, the obtained labels are mapped back to their corresponding positions in real space, facilitating visualization and interpretation.

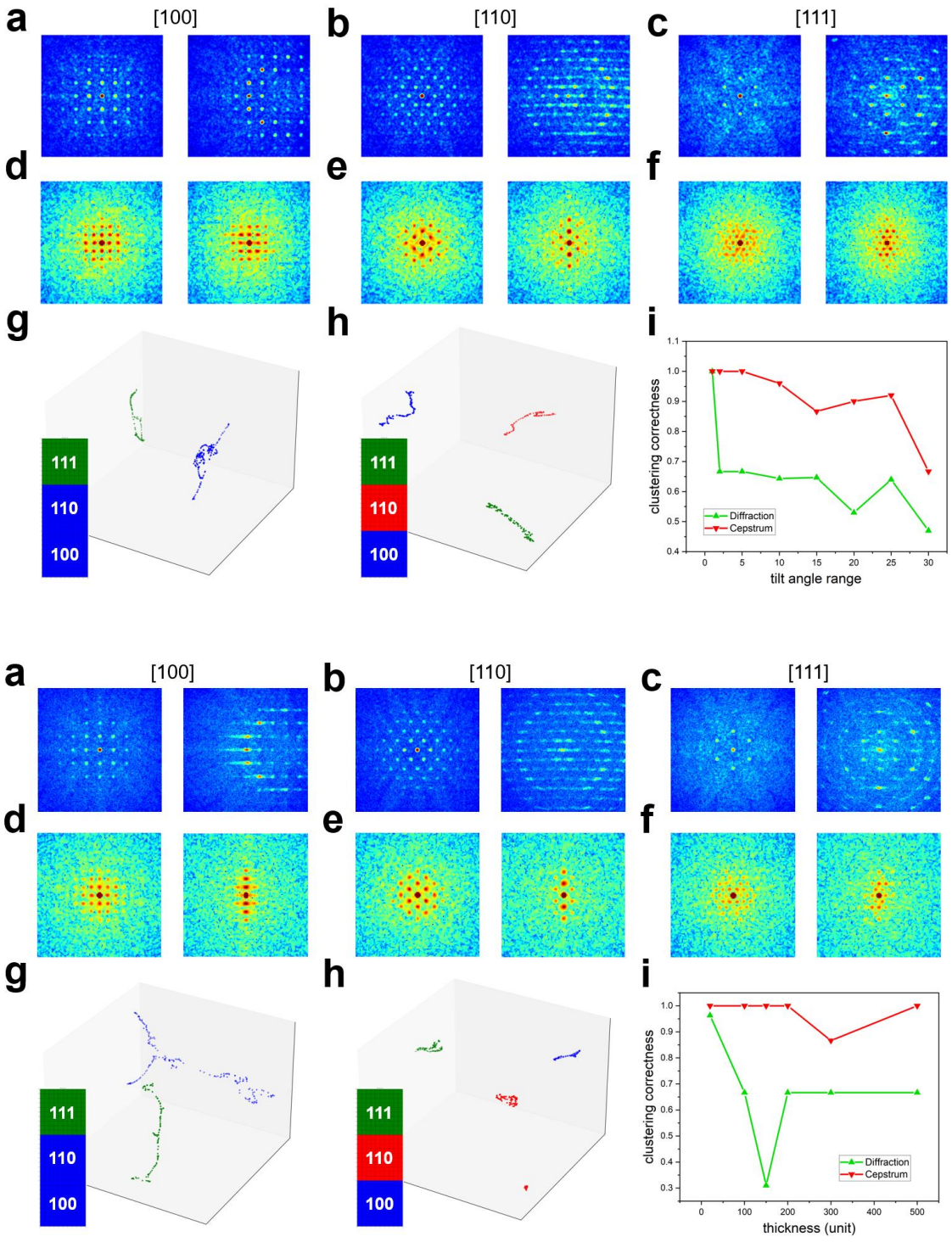

Figure 2| Example clustering results of a simulated dataset. (a-c) Diffraction images of [100], [110] and [111] orientation generated by the simulation program, along with their diffraction images tilted by 15° around the x-axis. **(d-f)** Cepstrum images of [100], [110] and [111] orientation generated by the simulation

program, along with their diffraction images tilted by 15° around the x-axis. **(g)** Results of the clustering algorithm based on the diffraction-generated manifold for the simulated dataset with a tilt range of (-5,+5)°. The manifold is continuous and hard to cluster. **(h)** Results of the clustering algorithm based on the cepstrum-generated manifold for the simulated dataset with a tilt range of (-5,+5)°. The manifold is separated into 3 individual clusters and is easily clusterable. **(i)** Line graphs depicting the accuracy of four different clustering algorithms applied to standard masked data/manifolds generated based on diffraction/cepstrum, with varying tilt angles in the simulated dataset.

The simulated dataset focuses on FCC structured Ag crystals, considering three different crystallographic orientations: [100], [110], and [111]. For each orientation, various x-axis rotation angles were applied, spanning 100 equally spaced values within predefined angle ranges (we choose (-1,+1), (-2,+2), (-5+5), (-10,+10), (-15,+15), (-20,+20), (-25,+25) and (-30,+30)°). Consequently, we obtained diffraction images under different crystallographic orientations and rotation angles, as illustrated in Fig. 2(a)-(c), which represent the [100], [110], and [111] orientations from left to right. Each pair of images includes the diffraction pattern without rotation (left) and the pattern obtained with a 1.5° rotation around the x-axis (right).

Subsequently, Gaussian filtering was applied to the diffraction patterns, followed by two-dimensional Fourier transformation to generate cepstrum images, presented in Fig. 2(d)-(f), corresponding to the previously mentioned orientations. Notably, introducing a minor rotation angle resulted in significant changes in the diffraction images compared to the absence of rotation. However, the cepstrum images showed comparatively minor variations. This is because rotation induces intensity changes in a specific area of the diffraction pattern, which corresponds to a corresponding change only in the intensity of the central 0th-order peak in the cepstrum image, with minimal impact on the intensity of other diffraction spots in the periphery.

The purpose of introducing rotation angles on different crystallographic orientations was to simulate local surface tilt, with the aim of investigating whether various clustering methods could eliminate the interference caused by tilt on the properties we intended to distinguish (in this case, the three distinct crystallographic orientations). To this end, we employed four different methods for clustering: based on diffraction or cepstrum 3D manifolds. Fig. 2(g) displays the clustering results for the simulated dataset with a tilt range of (-5° to +5°) based on the diffraction-generated manifold. It clearly shows that the manifold is separated into three distinct clusters and is easily separable. Conversely, Fig. 2(h) shows the clustering

results for the same dataset, but based on the diffraction-generated manifold, indicating a continuous manifold that is challenging to cluster.

Fig. 2(i) illustrates how the accuracy of the four clustering algorithms varies with the tilt angle range. Two key observations can be made: first, the accuracy of all four clustering algorithms decreases as the tilt angle range increases, as larger tilt angles introduce more significant disturbances to both diffraction and EWPC images, making it harder for the clustering algorithms to distinguish them from images of different crystallographic orientations. Second, the accuracy ranking of the four clustering algorithms is approximately as follows: cepstrum (manifold) > diffraction (manifold) > cepstrum (data) > diffraction (data). Notably, manifold-based clustering algorithms outperform those using only standard masked data (due to the cepstrum clustering algorithm's min sample size=50 setting, clustering algorithms based on diffraction standard masked data fail to produce any successful clusters). This finding aligns with the earlier conclusion drawn from Fig. 2(a)-(f), indicating that the use of cepstrum-based clustering can better eliminate the interference caused by tilt in diffraction images, resulting in higher accuracy.

The first experimental dataset under investigation pertains to the 2D ferroelectric material SnSe, characterized by intricate twin domain configurations. Notably, while the annular dark-field (ADF) image, as depicted in Figure 3a, reveals some observable domain contrast, it is imperative to acknowledge that the predominant contrast within this ADF image primarily arises from perturbations associated with surface tilting, a consequence of the inevitable sample transfer procedure. In Figure 3b, we present two distinct diffraction patterns originating from diverse tilted regions of the sample. When scrutinizing the disparities between these diffraction patterns, as illustrated in Figure 3c, it becomes evident that only two specific spots, demarcated within encircled regions, exhibit structural distinctions characteristic of the dipole configurations associated with twin domains. The remaining discrepancies within the diffraction patterns primarily arise from variations in intensity due to reciprocal lattice tilting. Consequently, it becomes apparent that the clustering outcomes derived from the diffraction patterns are chiefly influenced by the pronounced tilting characteristics (see Figure SI) rather than the intrinsic domain structures.

To mitigate the adverse effects stemming from sample tilting, we employ the EWPC transform on the diffraction data, subsequently presenting the corresponding cepstrum images in Figure 3d. It is noteworthy that within the framework of EWPC, the Fourier transform relies on the periodicity of the diffraction spots, wherein the contribution of individual diffraction spot intensities is relatively diminished compared to the spatial positioning of these spots. Consequently, the cepstrum images exhibit greater uniformity and a reduced susceptibility to distortion caused by lattice tilting in contrast to the original diffraction patterns. Subsequently, in Figure 3e, we observe that the discernible disparities among the cepstrum images

predominantly arise from the inherent structural distinctions embodied in the dipole configurations of twin domains, thereby endowing these distinctions with substantial physical significance.

The high-frequency segment of the cepstrum, delineated by the encircling ring in Figure 3e, is subjected to selection and subsequently embedded into the 3D manifold space, as illustrated in Figure 3f through the UMAP method. To ascertain the robustness of our findings, we employ both the K-Means and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering methods, facilitating a comparative analysis of their outcomes. Notably, the manifold space, as delineated by UMAP, already exhibits a substantial degree of data point separation, leading to a convergence of results between the two clustering techniques. In both instances, K-Means (refer to Figure SI) and HDBSCAN (see Figure 3g), the data points within the manifold space are categorically classified into four primary clusters, each of which corresponds to distinct domains within the real-space representation (see Figure SI and Figure 3h). The mean diffraction patterns originating from these four clusters, along with their variations, are meticulously presented in Figure SI. It is noteworthy that the parallel domains, represented by the color coding of yellow and red, as well as light and dark blue, collectively constitute the twin boundaries (as evident in Figure SI), while domains that are orthogonal to one another exhibit lattice rotations (as depicted in Figure SI).

Nevertheless, a salient disparity emerges between K-Means and HDBSCAN in their operational principles. HDBSCAN clustering is based on data point density, thereby designating data points situated distally from the cluster centers as outliers. Consequently, the manifold structure (**Fig. 3g**) features these outlier points prominently in black. Notably, with the exception of an anomaly in the upper-right corner, the outlier pixels within the real-space representation predominantly congregate along the boundaries of domains. This phenomenon is primarily attributed to the relatively large probe dimensions (~2.5 nm, detailed in the Methods section). Given that the twin boundaries are exceedingly thin, approximately one unit cell in thickness (~0.5 nm), the probe encompasses both sides of these boundaries, engendering distinct sets of diffraction spots that deviate from those present within single domains. These distinctive diffraction patterns are consequently identified as outliers.

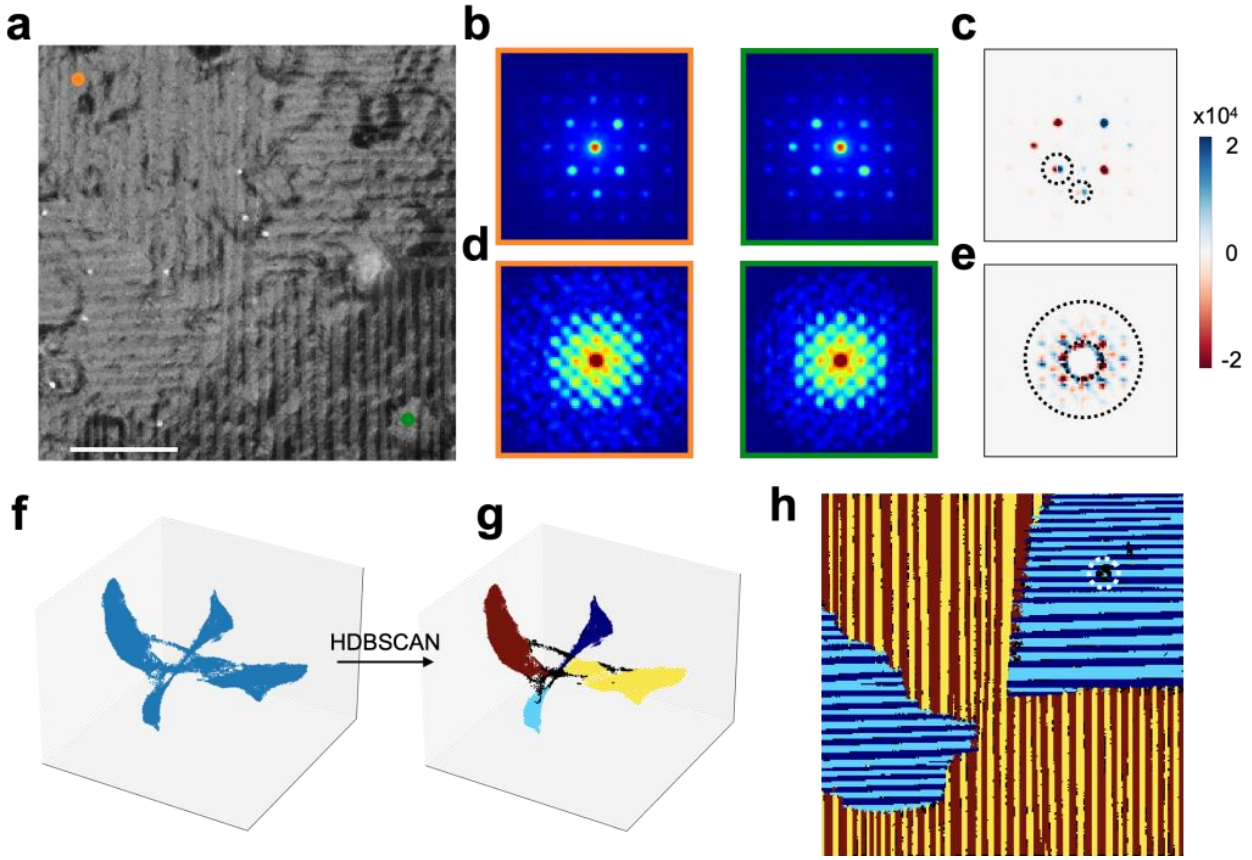

Figure 3| Clustering results of the 2D SnSe thin film. (a) The ADF image of the SnSe thin film where the domain contrast is coupled with tilt and surface variation. **(b)** Diffraction patterns corresponding to two points at top left (orange) and bottom right (green) in (a). **(c)** Differences of the two diffraction patterns. **(d)** Cepstrum images corresponding to the two diffraction patterns in (b). **(d)** Differences of the two cepstrum images, where the selected region is labeled in the black dot ring. **(e)** The 3D manifold structure of the cepstrum images. **(f)** The clustering results of the manifold structure through HDBSCAN method. Four distinct clusters are identified and color-coded, with central points designated as outliers and labeled in black. **(g)** Real space label map with four different domains. The large defect is labeled in white dot circle.

In addition to its application in two-dimensional (2D) thin materials, manifold learning in the cepstrum space has demonstrated its potential to enhance clustering performance even in thicker materials. Figure 4a presents an ADF (Annular Dark-Field) image depicting an Al-AlOx nanoparticle with a thickness extending to tens of nanometers. Notably, the surface exhibits roughness attributed to the rigorous high-temperature treatment applied. Consequently, the sample exhibits a near-alignment with the zone axis, leading to observable tilting variations across distinct regions within the diffraction patterns, as evident in Figure 4b.

These variations subsequently result in intensity fluctuations at the center of the diffraction patterns, as depicted in Figure 4c. When the clustering methods are directly applied to these diffraction patterns, the outcomes become contingent upon the fortuitous surface variations (see Figure SI) rather than encapsulating meaningful structural disparities. Therefore we target to decouple the tilting features and lattice features in the diffraction patterns in the frequency space.

According to the strong phase approximation model, an excitation error arises when the Ewald sphere intersects with the reciprocal lattice of finite thickness, giving rise to a low-frequency envelope function subject to shifting by varying tilting angles. Figure SI presents the low-frequency representations of the two diffraction patterns from Figure 4b, whereas their discrepancies are delineated in Figure 4e, highlighting the perceptible shift in the envelope functions. Notably, this low-frequency differential information within the diffraction data remains entangled with the lattice characteristics and proves challenging to isolate, as illustrated in Figure 4c. However, through the application of the EWPC method, the low-frequency features are effectively concentrated within the central regions of cepstrum images, enabling their separation. Consequently, only the high-frequency spots are retained for subsequent analytical processes, as depicted in Figure 4f.

The K-means clustering results in real space and manifold space are shown in Fig. 4g and 4h and the mean cepstrum images of each cluster are shown in Fig. 4i. Six clusters are identified on this Al-AlO nanoparticle. Two twin domains (No. 1, 2) situated along the [110] zone axis are discerned within the Al core, manifesting themselves as two spherical clusters in the manifold space. Four distinct AlOx clusters have been resolved, characterized by varying lattice orientations and degrees of crystallinity. These clusters encompass two single crystalline regions with dissimilar lattice orientations (No. 3, 4), a polycrystalline region (No. 5), and an amorphous region (No. 6).

While we also conducted clustering using the HDBSCAN method, as depicted in Figure SI, it is noteworthy that this approach yields results akin to those obtained through K-means clustering. However, it is important to acknowledge the presence of outlier points regardless of the chosen parameterization within the HDBSCAN method, as evidenced in Figure SI. Given the consistent and coherent outcomes obtained through the K-means clustering approach, we have elected to adopt the K-Means clustering labels as our final results for this specific Al-AlOx dataset. These discrete clusters serve as a foundation for conducting a more comprehensive investigation into the relationship between the lattice structure and the surface chemistry of the AlOx shell encompassing the Al-AlOx nanoparticles.[cite Aaron's paper].

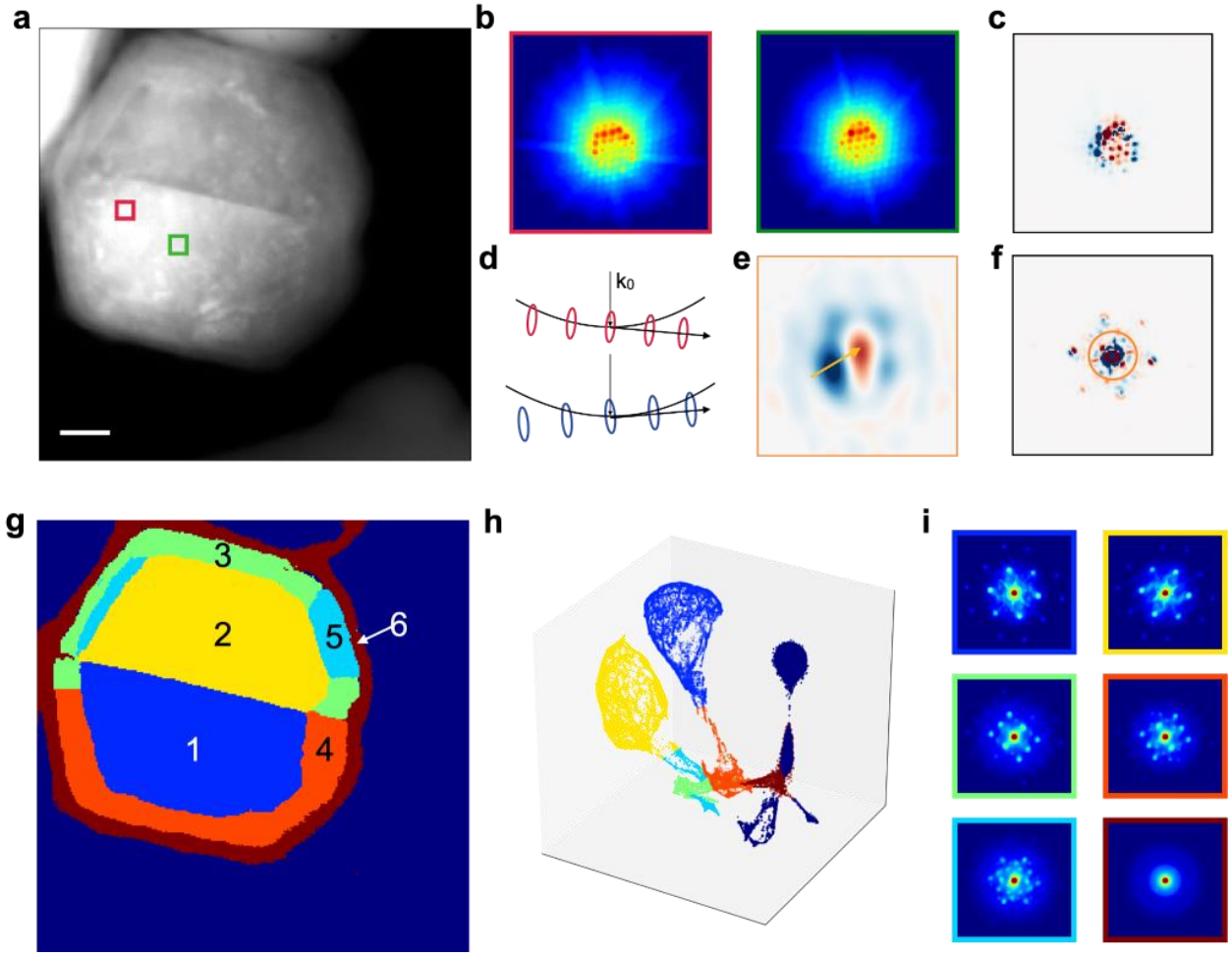

Figure 4| Clustering results of the Al-AlOx nano particle. (a) The ADF image of the nano particle. **(b)** Diffraction patterns corresponding to two points in (a). **(c)** Differences of the two diffraction patterns. **(d)** Schematic of the Eward sphere cutting the tilted reciprocal lattices with finite height. **(e)** Low frequency differences between the two diffraction patterns in (b), which shows the shift of the envelope function caused by lattice mistilt. **(f)** Differences of the two cepstrum images correspond to the two diffraction patterns in (b), the low frequency region is labeled in the ring. **(g-h)** K-Means clustering results in real space (g) and manifold space (f). **(i)** Mean cepstrum image of each cluster on the Al-AlOx nano particle.

In summary, our study introduces an effective data preprocessing technique that combines cepstrum analysis and manifold learning to enhance unsupervised learning performance on nanobeam 4D-STEM data. The EWPC method isolates high-frequency lattice signals from low-frequency noise caused by lattice tilt in cepstrum images. Meanwhile, the UMAP method projects lattice features into a lower-dimensional

manifold space based on similarities, facilitating data visualization and expediting clustering. We validate this approach on simulated data with manually defined ground truth and further apply it to real-world 2D SnSe and thicker Al-AlOx nanoparticles. Notably, our method accurately identifies novel material structures, including strain variations, lattice orientations, and crystallinity levels, while mitigating the influence of lattice mistilt. We anticipate that this unsupervised learning workflow, encompassing data transformation and clustering, will enhance the interpretation of intricate 4D-STEM datasets and contribute to the discovery and design of novel materials.