Crystallographic variant mapping using precession electron diffraction data

Marcus H. Hansen[1], Ainiu L. Wang[1], Jiaqi Dong[1], Yuwei Zhang[2], Tejas Umale[1], Sarbajit Banerjee[1,3], Patrick Shamberger[1], Matt Pharr[2], Ibrahim Karaman[1], Kelvin Y. Xie[1]

[1]Department of Materials Science and Engineering, Texas A&M University, College Station, TX 77843, USA.
[2]Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA.
[3]Department of Chemistry, Texas A&M University, College Station, TX 77843, USA.

Correspondence to: Prof. Kelvin Y. Xie, Department of Materials Science and Engineering, Texas A&M University, College Station, TX, 77843, USA. E-mail: kelvin_xie@tamu.edu

Abstract

In this work, we developed three methods to map crystallographic variants of samples at the nanoscale by analyzing precession electron diffraction data using a high-temperature shape memory alloy and a VO_2 thin film on sapphire as the model systems. The three methods are (I) a user-selecting-reference pattern approach, (II) an algorithm-selecting-reference-pattern approach, and (III) a k-means approach. In the first two approaches, Euclidean distance, Cosine, and Structural Similarity (SSIM) algorithms were assessed for the diffraction pattern similarity quantification. We demonstrated that the Euclidean distance and SSIM methods outperform the Cosine algorithm. We further revealed that the random noise in the diffraction data can dramatically affect similarity quantification. Denoising processes could improve the crystallographic mapping quality. With the three methods mentioned above, we were able to map the crystallographic variants in different materials systems, thus enabling fast variant number quantification and clear variant distribution visualization. The advantages and disadvantages of each approach are also discussed. We expect these methods to benefit researchers who work on martensitic materials, in which the variant information is critical to understand their properties and functionalities.

Keywords: Crystallographic variant mapping, precession electron diffraction (PED), image similarity quantification, k-means

INTRODUCTION

The martensitic transformation is an important phenomenon, which is widely observed in many metals [e.g., steels, shape memory alloys (SMAs), *etc.*][1-5] and ceramics (e.g., ZrO_2, VO_2, *etc.*)[6-8]. During a martensitic transformation, the martensite grains exhibit a well-defined crystallographic orientation relationship with the parent austenite grains. For example, in $Ni_{36.5}Ti_{48.5}Hf_{15}$ SMAs, the orientation relationship between the B2 austenite and B19′ martensite is described as [100]B2//[100]B19′ and (001)B2//(011)B19′[9]. Moreover, although the number of martensite grains in the same prior austenite grain can be numerous, the martensite variant number is limited. Theoretical predictions based on the lattice correspondences of B2 and B19 showed that 12 martensite variants are permitted[10]. Note that not all 12 martensite variants will show up due to self-accommodation constraints[10]. For example, only four martensite variants were observed in a solution-treated $Ni_{50.3}Ti_{29.7}Hf_{20}$ SMA, and there are well-defined orientation relationships between the martensite variants[11].

The transformation behavior and properties of solid-state phase change materials are dictated by their structure and substructure (e.g., martensite size, variant numbers, orientation relationship, the presence/absence of internal twins, *etc.*). However, martensite characterization, in particular identifying the martensite variant number and orientation relationship, has posed a challenge using conventional microscopy techniques. Traditional transmission electron microscopy (TEM) imaging can reveal the size of martensite grains but provides no information on the number of martensite variants and their distribution[12,13]. Electron backscatter diffraction (EBSD) can be used to map martensite variants, but the orientation maps are generally very noisy, with low indexing rates[14]. Because of its limited resolution, EBSD also fails to capture the crystallographic information if the martensite grains are small (e.g., plates or laths < 150 nm in width)[14]. Hence, there is a demand for techniques that can efficiently reveal martensite variant information with high spatial resolution.

Precession electron diffraction (PED) is a powerful characterization technique to reveal the crystal structure and orientation information at the nanoscale[15-22]. The electron beam in TEM is converged to a small probe (~ 1-3 nm) and rastered on the specimen. Precession (typically 0.3-0.8°) is applied to excite higher-order reflections and to reduce the dynamical effect[15,23]. The experimentally acquired diffraction patterns from each pixel are compared to the simulated diffraction patterns in a database to determine the crystal structure and orientation. The information is then used to create phase (crystal structure) and orientation maps. Naturally, the PED should be the ideal technique to obtain the martensite variant information in various materials. Unfortunately, there are several associated challenges. First, many martensite grains contain high-density twins, which leads to additional spots (and streaking) in diffraction patterns[11]. The simulated diffraction patterns in the database assume single crystals. Thus, the experimentally acquired diffraction patterns that contain both diffraction spots from the matrix and the twin cannot find a good match in the database and are considered bad indexing. Second, orientation indexing is generally poor for low-symmetry crystals (monoclinic, triclinic, *etc.*) in PED because the small differences between different sets of lattice planes cannot be easily distinguished by the orientation indexing software. The confusion leads to noisy orientation maps (an example, see Figure 1).

In this work, by taking advantage of the superior spatial resolution and the diffraction information from each pixel in PED, we developed three methods to create "crystallographic variant maps". All methods identify the crystallographic variants by comparing the diffraction patterns in the PED data, with the first approach more manual and the second and third more automatic. The advantages and disadvantages of each approach are described. Different methods to quantify the similarity between diffraction patterns, as well as their influence on the final crystallographic variant maps, are also discussed. These new semi-
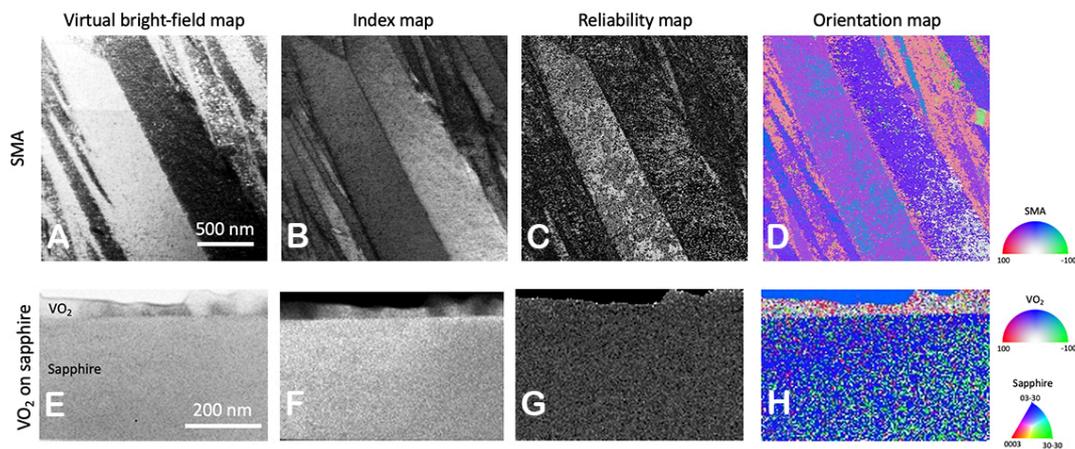
Figure 1. (A) VBF, (B) index, (C) reliability, and (D) orientation maps of the SMA sample. (E) VBF, (F) index, (G) reliability, and (H) orientation maps of the VO_2 on sapphire sample. Note the reliability and orientation maps are noisy in both samples.

automatic crystallographic mapping methods are open-source, easy to use, and less prone to human errors and can be useful tools for researchers who work on martensitic material systems.

MATERIALS AND METHODS

The two model systems we use in this study to develop the crystallographic variant mapping methods are (1) a $Ni_{50.3}Ti_{29.7}Hf_{20}$ SMA and (2) a VO_2 thin film deposited on a *c*-cut monocrystalline sapphire substrate. Both samples are martensitic at room temperature. The $Ni_{50.3}Ti_{29.7}Hf_{20}$ SMA has a monoclinic crystal structure, with a = 3.052 Å, b = 4.089 Å, c = 4.887 Å, and β= 102.91°. The VO_2 also has a monoclnic crystal structure, with a = 5.752 Å, b = 4.538 Å, c = 5.383 Å, and β = 122.65°[12]. The *b*-axis of VO_2 is parallel to the *c*-axis of sapphire (hexagonal crystal structure, a = 4.760 Å and c = 12.994 Å)[24]. The detailed processing information of these samples can be found in our previous work[11,25]. SMAs are known for their shape memory effect and super-elastic behavior, which make them widely used in mechanical, aerospace, and biomedical applications[4,26]. VO_2 undergoes an insulator-to-metal transition when heated up above the phase transformation temperature and is a promising material for neuromorphic computing[27,28]. The SMA sample was aged at 550 °C for 10 h to grow nanoscale precipitates. The diffraction patterns were obtained with a 200 kV *FEI Tecnai F20 ST* TEM equipped with a NanoMEGAS *ASTAR* system with a *Stingray* CCD camera to acquire the PED data. The diffraction pattern acquisition rate is approximately 0.06 s per pixel. The precession angle is 0.3°, and the electron beam spot size is approximately 3 nm. The 0.3° precession angle was selected because it offers a good combination of reduction in the dynamical effect and retaining a decent beam spot size. While larger precession angles can further reduce the dynamical effect, they result in larger beam spot sizes and, consequently, poorer spatial resolution for mapping. The SMA map is 285 × 285 pixels. The diffraction patterns were acquired with a 144 × 144 pixel resolution. The VO_2 on the sapphire map is 150 × 100 pixels. The diffraction patterns were acquired with a 580 × 580 pixel resolution to evaluate the effect of noise and denoise on the final crystallographic variant maps. The denoise was achieved by applying a 5-pixel radius Gaussian filter.

For the PED data reading and then processing, the ".blo" file is loaded and read using Hyperspy, which converts the file into a usable 4D array[29]. Hyperspy is an open-source Python library that provides tools to analyze multi-dimensional data, such as a function to read ".blo" files.

Regarding crystallographic variant mapping, three methods were developed. Both the first and second methods rely on comparing the diffraction patterns of each pixel to the reference patterns from the same PED dataset. The first method is more manual in that it requires the user to first choose the reference patterns. For each diffraction pattern in the dataset, the highest similarity value between all the reference patterns is used to determine which crystallographic variant it is most similar to. A crystallographic variant map is then generated based on the highest similarity values at each location in the data. The second method is more automatic, in which the algorithm will create new references when the similarity values between the test pattern and the existing reference patterns are lower than a threshold. In the first two methods, the Euclidean distance, Cosine, and Structural Similarity (SSIM) algorithms were used to quantify the similarity values between the diffraction patterns of each pixel to references.

The Euclidean distance algorithm[30] flattens the 2D image arrays to create 1D number arrays. The 1D number arrays can be viewed as vectors. For 144 × 144 pixel resolution diffraction patterns, the i value is 20,736 (i.e., 144^2). Hence the 1D number arrays correspond to vectors in 20,736 dimensions. The Euclidean distance, $d(u, v)$, between the test diffraction pattern and the reference pattern is calculated as the distance between the two vectors in 20,736 dimensions, which is expressed as $d(u, v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$, where u and v are intensity values of each pixel in the test and reference diffraction patterns, respectively. In PED, the diffraction patterns are stored as 8-bit images with the values of u and v ranging from 0 to 255. To turn the Euclidean distances into similarity values that range from 0 to 1, we describe Euclidean similarity ($S_{Euclidean}$) as $S_{Euclidean} = 1 - \frac{d(u,v)}{|u|+|v|}$, where $|u|$ and $|v|$ are the magnitudes of the vectors corresponding to the test and reference patterns, respectively. The magnitude is calculated as $|u| = \sqrt{\sum_{i}^{n} u_i^2}$. The same applies to v in the reference pattern. A $S_{Euclidean}$ value of 1 indicates that the test and reference diffraction patterns are identical. For diffraction patterns acquired with 580 × 580 pixel resolution, the i value is 336,400 (i.e., 580^2), and the vectors are 336,400 dimensional. The Euclidean distances and the similarities are calculated in the same way as described above.

The Cosine image comparison method[31] is similar to the Euclidean distance algorithm. The 2D image arrays are also transformed into 1D arrays (i.e., vectors at high dimensions). The Cosine method uses the angle between the two high-dimensional vectors describing the test and reference patterns. If the cosine angle between the arrays is low, the two images are similar. The Cosine similarity (S_{Cosine}) is calculated as $S_{Cosine} = 1 - \left|\frac{u \cdot v}{|u||v|}\right|$, where $u \cdot v$ is the dot product between the test (u) and reference (v) diffraction patterns. The S_{Cosine} value also ranges between 0 and 1, and a value of 1 means that the two images are the same and the angle between the corresponding two vectors is 0.

The SSIM image comparison method measures the structural similarity between two images (the reference and test diffraction patterns in this work), taking into account luminance, contrast, and structure[32]. Luminance (l) measures the brightness difference between the two images and is defined as $l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$, where μ_x and μ_y are the pixel intensity means of two images x and y, and c_1 is a constant. Contrast (c) captures the variation of brightness in the images and is defined as $c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$, where σ_x and σ_y are the pixel intensity variances of two images x and y, and c_2 is also a constant. The structure (s) captures the patterns and texture in the image and is defined as $s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$, where σ_{xy} is the pixel intensity covariance of images x and y, and c_3 is also a constant. The SSIM similarity combines the above measurements and is defined as $SSIM(x,y) = l(x,y) \cdot c(x,y) \cdot s(x,y)$. Similar to $S_{Euclidean}$ and S_{Cosine}, $SSIM(x,y)$ also ranges from 0 to 1, and

a value of 1 indicates that the two compared images are identical.

In the third crystallographic variant mapping method, the k-means method[33] was used to cluster the diffraction patterns into different groups, with each group representing one crystallographic variant. The algorithm uses the *VGG16 Keras* model (a pre-trained deep learning model that consists of 16 convolutional layers and three fully connected layers used for feature extraction) to reduce all diffraction patterns in a dataset into feature vectors[34]. The feature vectors capture key characteristics of each diffraction pattern, including structure and orientation. Next, the feature vectors are reduced from 4,096 to 100 components, each using kernel principal component analysis (kernel PCA) to lower the amount of data processed in the next step while retaining key information. Kernel PCA is a dimensionality reduction technique that uses kernel methods to transform nonlinear data[35]. These vectors are input into a k-means++ algorithm (an extension of conventional k-means but with much better with centroid initialization)[36]. The algorithm then sorts the vectors into k clusters, which are mapped to create a similarity map. For each k-value, the inertia (the sum of the squared distances of samples to their closest cluster centroid) is also calculated and shown in an "elbow graph" that can be used to manually determine an optimal k-value.

RESULTS AND DISCUSSION

Conventional PED characterization of the model systems

Figure 1 shows the virtual bright-field (VBF), index, reliability, and orientation maps of the SMA and VO_2 on sapphire, which are the model systems of this study. The VBF images are formed using the intensity of the direct beam in diffraction patterns and are similar to the conventional bright-field TEM micrographs but with less dynamical effect due to the beam precession. The index map is similar to the band contrast map in EBSD, in which brighter pixels indicate better matches between the experimentally acquired diffraction pattern and the simulated pattern. The reliability map is obtained by calculating the ratio of similarities between the best and second-best matches obtained through template matching. Brighter pixels in the reliability map indicate a higher level of confidence and reduced ambiguity between the best and second-best matches in orientation indexing. The concept of "reliability" will serve as a metric to assess the quality of our crystallographic mapping in subsequent sections of this work.

The SMA sample (B19', monoclinic crystal structure) shows a typical martensitic microstructure where the martensite grains are plate-like [Figure 1A]. The martensite plate thickness varies from tens to hundreds of nanometers. The corresponding index, reliability, and orientation maps [Figure 1B-D] show that multiple martensite variants exist in the sample, but the result is noisy. A VO_2 (monoclinic crystal structure at room temperature) thin film grows epitaxially on c-cut sapphire[37], which exhibits three crystallographic variants[38]. The VBF in Figure 1E shows that the VO_2 film is approximately 50-100 nm thick. The corresponding index, reliability, and orientation maps [Figure 1F-H] are extremely noisy, and the variants of VO_2 cannot be identified at all. The poor indexing of the martensitic SMA and VO_2 may be caused by the low symmetry nature of their monoclinic crystal structure, where the spacing and angles of different lattice planes are close to each other, leading to confusion when the software does the diffraction pattern indexing. Consequently, there is a demand for new methods to better illustrate the distribution of crystallographic variants in this wide group of materials.

Method 1: user-selecting-reference-pattern approach

In this method, the user will first peruse the dataset to identify how many crystallographic variants are present in the dataset and then select a reference pattern from each variant. We observed no apparent variation in the diffraction patterns among pixels within each crystallographic variant. This allows high flexibility in selecting the reference pattern for each variant by the user. Next, the diffraction pattern from

each pixel will be compared to all reference patterns, and the one with the highest similarity value will be used for the variant assignment of pixels. Each variant is represented by a color in the map. Before going into the detailed mapping results using the first method, we will briefly discuss similarity quantification using Euclidian distance, Cosine, and SSIM algorithms.

Figure 2 shows the VBF of the SMA sample (same as Figure 1A but with some brightness and contrast adjustment) and diffraction patterns from various martensite grains. One of the diffraction patterns was selected as the reference, whereas the others were compared to the reference. By visual inspection, it can be seen that the *Ref* and *B* patterns are from the same crystallographic variant, while each of the other patterns represents unique variants. Hence, we expect *Ref* and *B* patterns to display the highest similarity. The similarity results, calculated by Euclidian distance, Cosine, and SSIM algorithms, are shown in Table 1. Each comparison method returns a normalized similarity value ranging from 0 to 1. A 0 means the compared images are completely dissimilar, and a 1 means that they are exactly the same. As expected, diffraction pattern *B* has the highest similarity with the reference pattern for all three similarity methods. A major difference between the different methods is the overall spread in the similarity values between the patterns. Euclidean has a similarity value range of 0.081, Cosine has a small range of 0.026, and SSIM has the largest range of 0.137. We further examine the similarity comparison quality by calculating the reliability values. The reliability is calculated by dividing the highest similarity by the second highest similarity. Larger values indicate higher reliability. The reliability of the Euclidean, Cosine, and SSIM methods are 1.044, 1.006, and 1.083, respectively. The small range and low reliability for the Cosine algorithm may lead to inconclusive results when the two diffraction patterns are similar, which will be demonstrated shortly.

Figure 3 shows the crystallographic variant maps generated via the Euclidean distance [Figure 3A], Cosine [Figure 3B], and SSIM [Figure 3C] algorithms, along with a manually drawn map [Figure 3D]. The manually drawn map is created by a person inspecting the dataset to determine the variant regions and is treated as the baseline. Both the Euclidean and SSIM maps are exceptionally close to the manually drawn map. The four predominant variants (red, yellow, green, and cyan in color) and two minor variants (blue and magenta in color) are clearly revealed. Note that the Euclidean map is a bit noisy in the upper right region, while the SSIM maps provide a cleaner map. Unfortunately, the Cosine algorithm produced poor results. The map is noisy, and the large cyan variant in the center is split into three separate variants. This may be due to the varying diffraction pattern intensity along the variant. When generating the crystallographic maps using our methods [Figure 3A-C], the colors are selected on a hue saturation value (HSV) color wheel and spaced equally apart based on the number of reference points to distinguish them from each other. The colors were adjusted manually for clarity as needed. Generating the above maps [Figure 3A-C] only took tens of seconds or a few minutes using a computer with an Intel i7 13700K CPU and Nvidia RTX 3080 GPU, but it took a student several hours to generate the manually drawn map [Figure 3D].

To investigate the effect of noise of the diffraction patterns on the similarity quantification and the resultant crystallographic orientation maps, we used the VO_2 thin film deposited on a *c*-cut monocrystalline sapphire. The diffraction patterns were acquired with 580 × 580 resolution. The 144 × 144 diffraction pattern resolution in the previous SMA example was a result of binning the 580 × 580 original data by the NanoMEGAS commercial software during the data acquisition. Hence, the 580 × 580 resolution diffraction patterns in this case study are much noisier. Figure 4 shows the VBF of the VO_2 thin film on the sapphire substrate with diffraction patterns taken from different pixels. A diffraction pattern from the sapphire substrate was selected as the reference pattern. The diffraction patterns from pixels A, B, C, and D are from sapphire, VO_2 variant 1, VO_2 variant 2, and vacuum, respectively. Since diffraction pattern A was also taken

Table 1. Euclidean distance, Cosine, and SSIM similarity values compared to the reference pattern, along with the reliability values, in the SMA sample

Diffraction pattern	Similarity method		
	Euclidean	Cosine	SSIM
A	0.904	0.971	0.834
B	0.964	0.997	0.903
C	0.918	0.986	0.777
D	0.883	0.986	0.766
E	0.923	0.991	0.820
Reliability	1.044	1.006	1.083
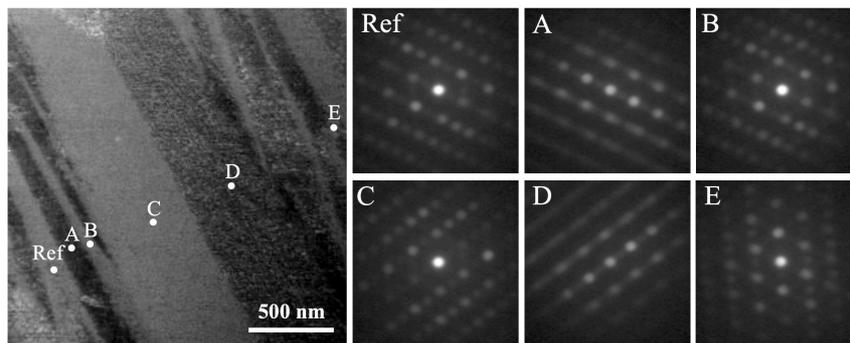
Figure 2. VBF image of the SMA sample (left) and the selected diffraction patterns (right) for similarity quantification.
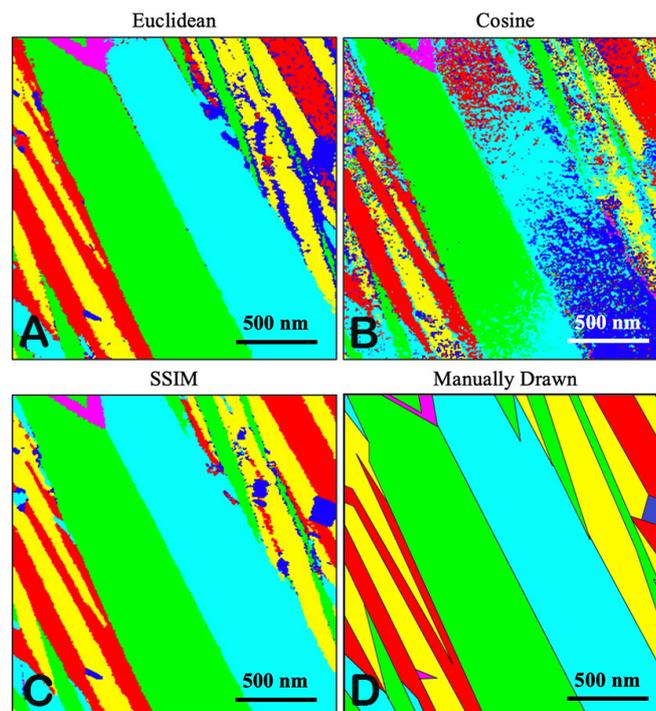
Figure 3. Crystallographic variant maps of the SMA sample generated using the (A) Euclidean distance, (B) Cosine, and (C) SSIM algorithms. (D) A mannually drawn crystallographic variant map of the same area serving as a baseline.
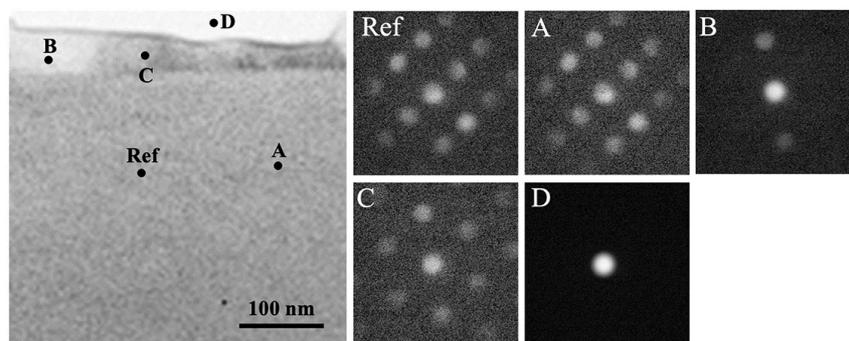
Figure 4. VBF image of VO_2 thin film on sapphire (left) and the selected diffraction patterns (right) for similarity quantification.

from the sapphire substrate, it is expected to have the highest similarity with the reference pattern. Diffraction pattern D was taken from the vacuum, in which there is only a direct beam without a diffracted beam; it is expected to have the least similarity with the reference pattern. Table 2 summarizes similarity values calculated between the selected diffraction patterns compared to the reference pattern using the Euclidean distance, Cosine, and SSIM algorithms. The Euclidean similarity shows the highest value for A and the lowest for D, as expected. The Cosine algorithm produced a similar distribution as the highest value for A and lowest for D while switching the order for B and C. In contrast, the SSIM method comparison had low overall similarity values and ranked pattern B higher than A. We also note that all the similarity values calculated here are generally lower than the binned case (the SMA example). This observation has two important implications. First, the presence of noise decreases the similarity values, although the two diffraction patterns are visually similar. Second, the presence of noise, though not significant to human eyes, can assign higher similarity values even to the visually dissimilar diffraction patterns.

As expected, the generated similarity maps generally display poor quality [Figure 5]. The Euclidean distance map [Figure 5A] is the best out of the three algorithms. The vacuum (black), sapphire substrate (red), and two variants of VO_2 (blue and green) were identified, but small parts VO_2 thin film were wrongly identified as the sapphire substrate. Note that there should be three variants of VO_2 on *c*-cut sapphire. Only two were captured in the figure due to the small mapping area. In the Cosine map [Figure 5B], large parts of the VO_2 thin film were mistakably identified as the sapphire substrate. The SSIM algorithm [Figure 5C] performed the worst, in which the film and substrate were not distinguished, albeit the two different variants in the VO_2 thin film. These observations suggest that the Euclidean distance method is most noise-tolerant when computing the similarity between two images. The accuracy of the SSIM method can be dramatically affected by noise in the diffraction patterns.

Removing the noise can dramatically improve the accuracy of diffraction pattern similarity quantification, which subsequently helps generate much improved crystallographic variant maps. One common way to denoise is by binning. For example, the NanoMEGAS commercial software can do binning that reduces the diffraction patterns from 580 × 580 to 144 × 144 resolution (an example see the SMA example in this work). Another common method is to apply a Gaussian filter[39]. Here, we explore the effect of denoise on the similarity quantification and crystallographic variant mapping accuracy using a Gaussian filter. Figure 6 shows the data similar to Figure 4, but a Gaussian filter (radius = 5) was applied to all diffraction patterns in Figure 6. With the Gaussian filter, the overall noise in each diffraction pattern is reduced. The Gaussian filter improved the similarity values for all methods by increasing the range and more accurately identifying the most similar diffraction patterns [Table 3]. The Euclidean similarity between patterns A and Ref is 0.910 after filtering compared to 0.707 using the raw data. The filtered diffraction patterns from the two VO_2

Table 2. Euclidean, Cosine, and SSIM similarity values compared to the reference pattern, along with the reliability values, in the VO_2 on sapphire sample

Diffraction pattern	Similarity method		
	Euclidean	Cosine	SSIM
A	0.707	0.841	0.045
B	0.632	0.807	0.046
C	0.678	0.804	0.033
D	0.421	0.771	0.035
Reliability	1.043	1.042	1.022

Table 3. Euclidean, Cosine, and SSIM similarity values compared to the reference pattern, along with the reliability values, in the VO_2 on sapphire sample denoised with a Gaussian filter applied (radius = 5)

Diffraction pattern	Similarity method		
	Euclidean	Cosine	SSIM
A	0.910	0.919	0.846
B	0.600	0.907	0.666
C	0.723	0.903	0.718
D	0.412	0.284	0.276
Reliability	1.259	1.013	1.178
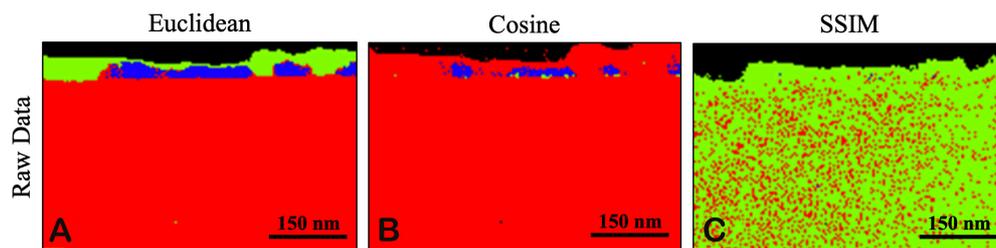
Figure 5. Crystallographic variant maps of the VO_2 thin film on sapphire sample generated using the (A) Euclidean distance, (B) Cosine, and (C) SSIM algorithms.
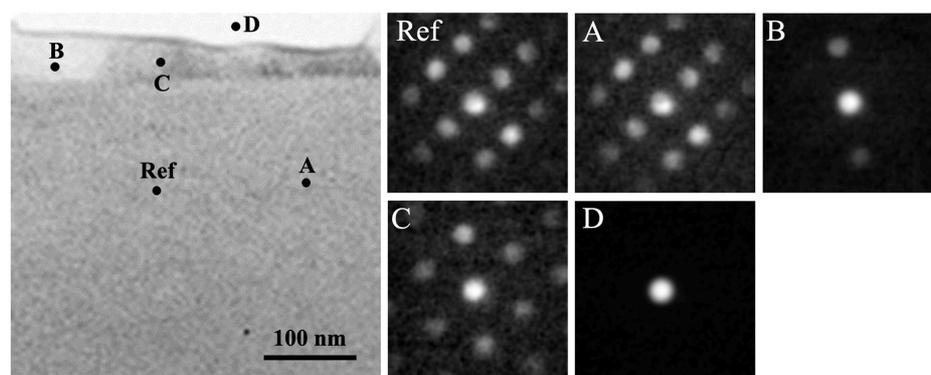
Figure 6. VBF image of VO_2 thin film on sapphire (left) and the selected diffraction patterns (right) after Gaussian filtering (radius = 5) for similarity quantification.

variants are very different from the Ref pattern (similarity values are 0.600 and 0.723). The difference is the largest when comparing the vacuum diffraction pattern with the Ref pattern, which yields a similarity value

of only 0.412. The Cosine method correctly identified that pattern A has the highest similarity value (0.919), but the ones from the two VO_2 variants also showed high similarity values (0.907 and 0.903), which may lead to crystallographic variant misidentification. Among the three algorithms, the Cosine method yields the lowest reliability value. This shortcoming of the Cosine method is consistent with the observation made in the SMA samples [Table 1]. The SSIM algorithm yields a similar trend to the Euclidean distance approach but shows low similarity values in absolute numbers. Taken together, reducing noise with a Gaussian filter significantly improved the similarity quantification across the three algorithms.

Reducing noise also improved the crystallographic variant mapping. Figure 7 shows the crystallographic variant maps generated using the Euclidean distance, Cosine, and SSIM algorithms. All methods produced good results where the sapphire substrate (red), VO_2 variants (blue and green), and vacuum (black) are clearly resolved. A manually drawn baseline map is not provided here because the interface between the two VO_2 variants is curved and difficult to draw. A careful examination of the diffraction patterns in this dataset confirmed that the maps shown in Figure 7 are indeed correct.

Method 2: algorithm-selecting-reference-pattern approach
The previous method, though powerful and efficient in generating the crystallographic variant maps, requires user input to select all known crystallographic variants. In some cases, the number and location of the crystallographic variants are unknown. Here, we describe a new method that is built upon the user-selecting-reference-pattern approach (Method 1) but uses the algorithm to select reference patterns to generate crystallographic variant maps.

This method starts at a pixel in the PED data (0, 0) by default or a pixel position chosen by the user. The diffraction pattern corresponding to the pixel is treated as the first reference and is added to a list that will contain reference patterns in the dataset. Each reference pattern in the list represents a unique crystallographic variant. The diffraction patterns from the following pixels will be compared to the reference pattern, and the similarity values (using Euclidean distance, Cosine, or SSIM) will be calculated. The starting pixel will be compared to all other pixels, and the minimum similarity value will be used as the initial cut-off value. If the calculated similarity value of the pixel is greater than the cut-off value, the algorithm treats the pixel the same as the reference pattern variant. If the calculated similarity value is less than the cut-off value, the algorithm will save the diffraction pattern of the pixel as a new reference. The diffraction patterns from the following pixels will be compared to the two references to determine which variant they belong to. This process iterates. If a pixel has one or more similarity values greater than the cut-off, the variant is determined by the one with the highest similarity. If a pixel has all similarity values less than the cut-off, the pixel itself will be added to the reference list and serve as a new variant. This process repeats for all pixels. Once all the pixels have been compared, the cut-off value increases by a step; by default, there are 20 steps between the initial minimum similarity and the maximum similarity values. The process repeats using the increased cut-off value and starting with the list of unique crystallographic variants from the previous iteration.

The above-mentioned mapping method was first used on the SMA dataset to produce a series of similarity maps at varying cut-off values. In the SMA dataset, the (0, 0) point is located at a boundary between two variants that may skew the variant identification. A user-selected point in the middle of a large variant, (116, 168), was used, and the corresponding diffraction pattern serves as the first reference pattern. Figure shows the SSIM similarity maps with various cut-off values ranging from 0.705 to 0.787. The optional initial point and comparison method are the only user inputs using this method. Note the small cut-off range generates very different crystallographic maps. With low cut-off values, such as 0.705, the four predominant
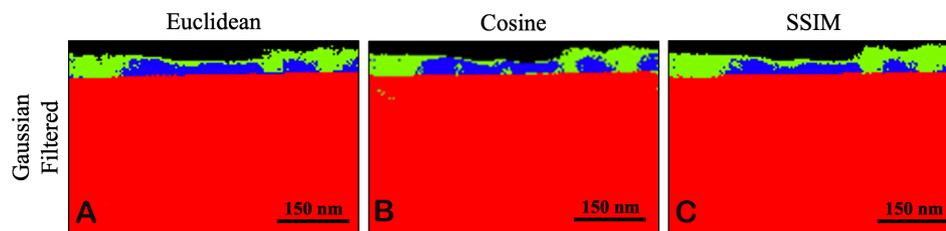
Figure 7. Crystallographic variant maps of the VO$_2$ thin film on sapphire sample generated using the (A) Euclidean distance, (B) Cosine, and (C) SSIM algorithms. The diffraction patterns in the PED data were filtered with a Gaussian filter (radius = 5) before generating the maps.

variants merged into only two "variants". With increasing cut-off values, new variants emerge [Figure 8A]. With the cut-off value at 0.769 [Figure 8D], the crystallographic variant map agrees very well with the baseline map in Figure 3D, where both maps contain four predominant and two minor variants [Figure 8D]. Further Increasing the similarity threshold splits variants into two or more pseudo-variants that are differentiated by small differences. For example, with the cut-off value of 0.787, a new dark green variant was generated from the yellow variant, and the magenta variant split from the cyan variant [Figure 8F]. The analysis and crystallographic variant mapping of this dataset took approximately 1 hour to complete.

This algorithm was also applied to the Gaussian-filtered VO$_2$ using the Euclidean similarity method. For this sample, the point (22, 13) in one of the VO$_2$ variants was selected as the initial point. Figure 9 shows the generated crystallographic variant maps with the cut-off increasing from 0.627 to 0.816. Again, such a relatively narrow range of cut-offs led to drastically different crystallographic variant maps. With a low cut-off value of 0.627, only the sample and vacuum were identified, but with low accuracy [Figure 9A]. Part of the VO$_2$ thin film was mistakably identified as vacuum. The sapphire substrate and two VO$_2$ variants were not resolved. The crystallographic variant identification improves as the cut-off value increases. At the cut-off value of 0.735, the most accurate map was generated [Figure 9C]. Further increasing the cut-off results in the variants splitting into pseudo-variants, as shown in Figure 9D and E. The analysis and crystallographic variant mapping of this dataset took approximately 20 min to complete.

The algorithm-selecting-reference-pattern approach can be a useful tool to accurately identify and map similar crystallographic variants in a dataset with minimal user input. Some user input is still needed to select the optional first point and to review the generated similarity maps to determine the ideal similarity cut-off threshold. However, this method is primarily automated and takes much less human effort than the user-selecting-reference-pattern approach (Method 1). Based on the observations made in the above two case studies, it can be seen that small changes in the cut-off value can result in large differences in the final crystallographic variant map output. Lower similarity cut-off values underestimate crystallographic variants, and higher similarity cut-off values overestimate them. User analysis is required to review the generated similarity maps to determine the optimal cut-off value for variant mapping. The computational time varies depending on the size of the dataset, the comparison method used, and the number of variants in the dataset. Typically, it is expected to take between 0.5 and 2 h to generate the similarity maps for a dataset.

Method 3: *k*-means approach
Unsupervised learning is widely used in image segmentation to group different regions in an image based on the pixel intensities and other features[21,22,40,41]. Martineau *et al.* and Bergh *et al.* successfully applied unsupervised learning techniques (e.g., non-negative matrix factorization and fuzzy clustering) to learn significant microstructural features in the PED data to achieve dimensionality reduction and overlapping
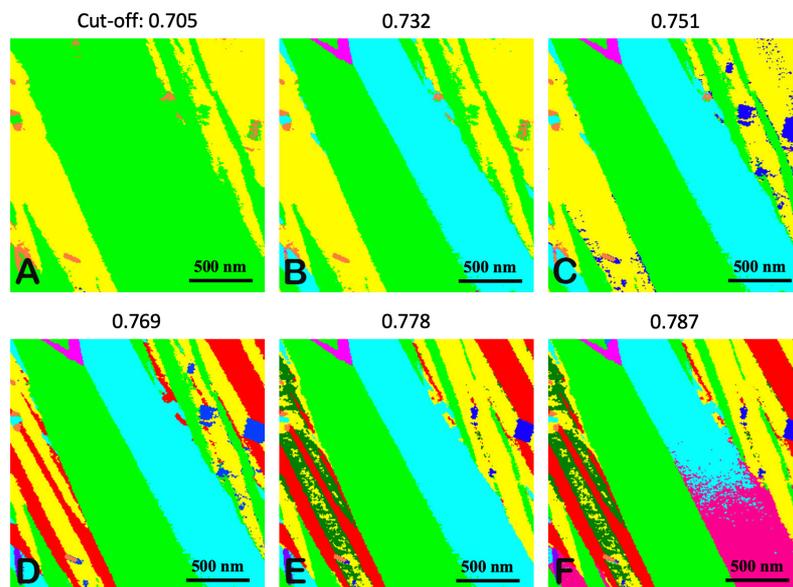
Figure 8. Crystallographic variant maps generated using the SMA dataset and the SSIM comparison method with (A) 0.705, (B) 0.732, (C) 0.751, (D) 0.769, (E) 0.778, and (F) 0.787 cut-off values.
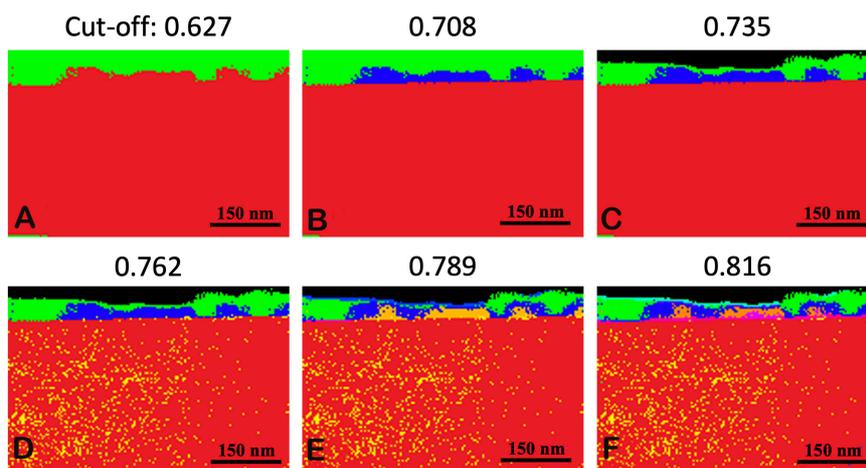
Figure 9. Crystallographic variant maps generated using the Gaussian-filtered VO_2 on sapphire dataset and the Euclidean comparison method with (A) 0.627, (B) 0.707, (C) 0.735, (D) 0.762, (E) 0.789, and (F) 0.816 cut-off values.

pattern unmixing[21,22]. Here, we employed k-means (a common unsupervised learning algorithm) in an approach that is different from the previous two reference pattern-based methods. The advantage of this unsupervised clustering algorithm lies in its ability to partition unlabeled data. After loading the data with *HyperSpy*, each diffraction pattern is processed into a feature vector using the *VGG16* model from *Keras*[34]. The dimensionality of each feature vector is then reduced with kernel PCA. In the user interface, the "elbow graph" is generated, allowing the user to determine an optimal k-value (the number of crystallographic variants in the data set) and input it into the k-means algorithm to create a crystallographic orientation map.

The elbow graph and the crystallographic variant maps for the SMA and Gaussian-filtered VO_2 on sapphire datasets with different k-values are shown in Figure 10. The SMA elbow graph [Figure 10A] has a rapidly dropping inertia from $k = 1$ to 3, after which the slope begins to flatten out. This indicates an optimal k-value of 3 or 4, as both the inertia and k at those points are minimized. From the previous discussion, the mapped region shows four predominant and two minor variants. Hence, the optimal k-value from k-means underestimates the variant numbers. Figure 10B shows crystallographic variant maps for $k = 3$, 4, and 5. In $k = 4$, the four predominant martensite variants were identified, but the two minor variants were lost. When k increased to 5, the minor variants were not revealed. Instead, a high level of noise was observed in the martensite plates. The k-means analysis took roughly 7 h on the SMA dataset. Similar observations were made in the Gaussian-filtered VO_2 on sapphire datasets. The elbow graph in Figure 10C indicates an optimal k-value of 2 or 3. From the previous discussion, the dataset should contain four crystallographic variants (sapphire, two VO_2 variants, and vacuum). Again, the optimal k-value tends to underestimate the variant numbers. Figure 10D shows crystallographic variant maps for $k = 3$, 4, and 5. In the map with $k = 3$, the sapphire substrate, VO_2 film, and vacuum were identified. However, part of the VO_2 film was erroneously marked as vacuum, as indicated by the arrows. In $k = 4$, the vacuum and thin film regions remain the same, but the sapphire substrate split into two pseudo-variants. In $k = 5$, the substrate region remains split, and the vacuum and thin film regions are accurately revealed, with the second VO_2 variant now showing. The k-means analysis took roughly 40 min on the cropped VO_2 on the sapphire dataset (100×100 pixels). Based on the above observations, the k-means method can provide general information on the crystallographic variants in the materials with minimal user input but can fail to capture minor variants and may falsely split variants into pseudo-variants.

It should be noted that while the k-means algorithm is well-suited for identifying similar-sized spherical clusters, it may not be the ideal choice for the SMA and VO_2 model systems. For example, in the case of the VO_2 on the sapphire dataset, sapphire, two variants of VO_2, and vacuum exhibit significant differences in data sizes, which can potentially lead to misidentifications when using $k = 3$ or 4. To address this challenge, alternative unsupervised machine learning techniques, such as density-based spatial clustering of applications with noise (DBSCAN), mean shift, and Gaussian mixture models (GMM), could be considered in future studies for more effective identification of crystallographic variants using the PED data[42]. Nevertheless, the use of k-means serves as a baseline for comparing the performance of other unsupervised learning techniques in future applications.

Method comparison

Each approach described above has ideal use cases along with advantages and disadvantages. The user-selecting-reference-pattern approach (Method 1) is best suited when the number and location of crystallographic variants in the sample are known. A region can be mislabeled if a reference point for the diffraction pattern is not selected. Inversely, if more than one reference point is selected for a pattern, a variant region can be divided between the selected points. Hence, if the user is not familiar with the material, this method can be prone to human errors. When all the variants are selected correctly, this approach can produce the most accurate similarity maps with the lowest computational cost (generally a few minutes). Among the diffraction pattern similarity quantification algorithms, both the Euclidean and SSIM methods outperform the Cosine method in generating more accurate crystallographic variant maps.

The algorithmic-selecting-reference pattern approach (Method 2) omits the need for the user to select the reference pattern of each variant. This approach can generate similarity maps automatically with only user input to select the ideal cut-off threshold based on the generated maps. The cut-off values in each map govern the new variant generation. Too high of a value will capture the small differences of diffraction patterns within the same variant, thus splitting one variant into two or more pseudo-variants. Too low of a
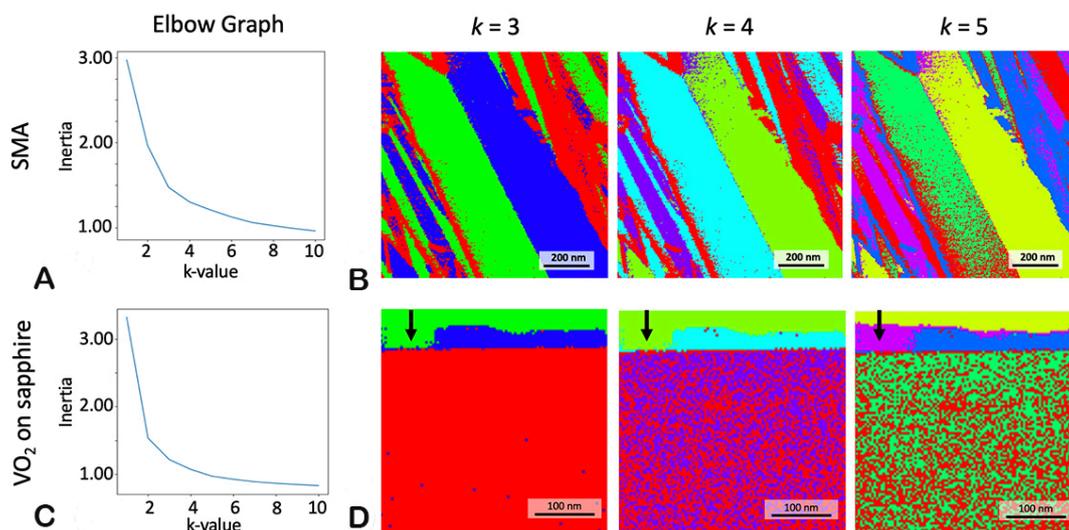
Figure 10. (A) Inertia *vs. k*-value elbow graph and (B) crystallographic variant maps with *k* = 3, *k* = 4, and *k* = 5 generated using the *k*-means clustering method on the SMA dataset. (C) Inertia vs. *k*-value elbow graph and (D) crystallographic variant maps with *k* = 3, *k* = 4, and *k* = 5 generated using the *k*-means clustering method on a portion of the Gaussian-filtered VO$_2$ on the sapphire dataset.

value will merge variants with different but similar diffraction patterns. Another challenge is that a small change of the cut-off value can lead to drastically different crystallographic variant maps. Generating the similarity maps with this method usually takes up to an hour on average.

The *k*-means approach (Method 3) can process the dataset and produce similarity maps automatically at various *k*-values. The only user input is the *k*-value (the crystallographic variant number of the mapped area). This number is more intuitive than the cut-off value in the previous method. Moreover, only one round of computation is required to generate the crystallographic orientation maps with various *k*-values. Changing the *k*-value and corresponding similarity map for a dataset is near instantaneous, allowing the user to quickly compare maps with varying numbers of diffraction pattern clusters. Unfortunately, the *k*-means approach tends to generate variant maps with the lowest quality. The generated similarity maps lack fine details, and noise and the number of pseudo-variants increase as the *k*-value increases. Additionally, the *k*-means approach is the slowest to compute, as it generally takes several hours to complete the analysis. This approach has room for improvement by experimenting with other unsupervised machine learning models.

These approaches can also be combined to help mitigate the shortcomings of a particular method. For example, in an unknown sample, the algorithm-assisted (Method 2) or the *k*-means method (Method 3) can be used to determine the approximate variant number and location, which can then be input into the user-assisted algorithm (Method 1), with a manual review, to create an accurate similarity map. It is also worth mentioning that although the methods developed in this work were designed to analyze PED datasets, the same approach can also be applied to the four-dimensional scanning transmission electron microscopy (4D-STEM) datasets, in which the diffraction patterns of each pixel are formed via convergent-beam electron diffraction[43], to determine the crystallographic variants.

It is also worth noting that creating TEM specimens with generally uniform thickness is imperative for all algorithms to correctly identify the crystallographic variants. Uneven specimen thickness can significantly impact the diffraction patterns of individual pixels. For example, within a single variant, regions with

greater thickness will exhibit decreased diffraction spot intensity and increased background noise[44]. Thick regions also limit the reduction of the dynamical effect, even with the application of precession[45]. These factors ultimately affect similarity comparisons and can result in poor crystallographic variant mapping.

CONCLUSIONS

In this work, we developed new crystallographic mapping methods to quantify the variant numbers and reveal the variant distribution by analyzing PED data. An SMA sample and a VO_2 thin film on sapphire sample were used as the model systems. Here are some key take-away messages.

Three methods were developed to map the crystallographic variants of materials using PED data. These methods are (I) a user-selecting-reference-pattern approach; (II) an algorithm-selecting-reference-pattern approach; and (III) a k-means approach.

When applying the first two approaches, Euclidean, Cosine, and SSIM were used to calculate the similarities between diffraction patterns. The Euclidean and SSIM algorithms outperform the Cosine method and can generate reliable similarity values for crystallographic variant mapping.

Background noise in diffraction patterns is detrimental to the similarity quantification. Denoising, either via binning or applying a filter (e.g., Gaussian filter), can dramatically improve the crystallographic variant mapping quality.

In the first approach, the user will select the reference patterns, and each reference represents a unique crystallographic variant. All diffraction patterns in the dataset will then be compared to the reference patterns. The one with the highest similarity will be used for assigning the crystallographic variant. This approach is fast and accurate when using the Euclidean and SSIM similarities, but it requires the user to have a good understanding of the sample to correctly select the reference patterns for accurate crystallographic variant map generation.

The second approach is similar to the first one, but the algorithm, instead of the user, selects the reference patterns. When scanning through the dataset, a new reference (representing a new variant) will be created whenever the new diffraction patterns are dissimilar, based on a cut-off value, to the reference patterns stored in the list. Hence, it is a dynamic process. This approach is moderately fast and can generate high-quality crystallographic variant maps. However, the map quality is sensitive to the cut-off value used.

In the third approach, the algorithm uses the k-means to cluster similar diffraction patterns into different categories (variants). The only input is the k-value (variant number). Unfortunately, this approach is computationally expensive, and the crystallographic variant maps are of low quality. Nevertheless, this approach can still provide some information on the general variant number and distribution, which can be helpful if the user is not familiar with the material system.

DECLARATIONS

Acknowledgments

Authors' contributions
Designed this study and performed data analysis and interpretation: Hansen MH, Wang AL, Xie KY
Prepared the samples and performed data acquisition: Dong J, Zhang Y, Umale T
All authors contributed to the intellectual merit and writing of the manuscript.

Availability of data and materials
The code used for our algorithms can be found on GitHub (https://github.com/TAMU-MHansen/PED-Similarity-Mapping). A tutorial on how to use our code can be found on YouTube: https://www.youtube.com/watch?v=Jr_QxVrVcbg.

Conflicts of interest
All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

REFERENCES
1. Nishiyama Z. Martensitic transformation. Amsterdam: Elsevier; 2012.
2. Bhadeshia H, Honeycombe R. Steels: microstructure and properties. Oxford: Butterworth-Heinemann; 2017.
3. Zaefferer S, Ohlert J, Bleck W. A study of microstructure, transformation mechanisms and correlation between microstructure and mechanical properties of a low alloyed TRIP steel. *Acta Mater* 2004;52:2765-78. DOI
4. Lagoudas DC. Shape memory alloys: modeling and engineering applications. Berlin: Springer; 2008.
5. Otsuka K, Ren X. Physical metallurgy of Ti-Ni-based shape memory alloys. *Prog Mater Sci* 2005;50:511-678. DOI
6. Garvie RC, Nicholson PS. Phase analysis in zirconia systems. *J Am Ceram Soc* 1972;55:303-5. DOI
7. Morin FJ. Oxides which show a metal-to-insulator transition at the neel temperature. *Phys Rev Lett* 1959;3:34-6. DOI
8. Goodenough JB. The two components of the crystallographic transition in VO_2. *J Solid State Chem* 1971;3:490-500. DOI
9. Han X, Zou W, Wang R, Zhang Z, Yang D. Structure and substructure of martensite in a $Ti_{36.5}Ni_{48.5}Hf_{15}$ high temperature shape memory alloy. *Acta Mater* 1996;44:3711-21. DOI
10. Nishida M, Nishiura T, Kawano H, Inamura T. Self-accommodation of B19′ martensite in Ti-Ni shape memory alloys - part I. morphological and crystallographic studies of the variant selection rule. *Philos Mag* 2012;92:2215-33. DOI
11. Dong J, Umale T, Young B, Karaman I, Xie KY. Structure and substructure characterization of solution-treated $Ni_{50.3}Ti_{29.7}Hf_{20}$ high-temperature shape memory alloy. *Scr Mater* 2022;219:114888. DOI
12. Evirgen A, Karaman I, Santamarta R, Pons J, Hayrettin C, Noebe R. Relationship between crystallographic compatibility and thermal hysteresis in Ni-rich NiTiHf and NiTiZr high temperature shape memory alloys. *Acta Mater* 2016;121:374-83. DOI
13. Karaca H, Saghaian S, Ded G, et al. Effects of nanoprecipitation on the shape memory and material properties of an Ni-rich NiTiHf high temperature shape memory alloy. *Acta Mater* 2013;61:7422-31. DOI
14. Cayron C. What EBSD and TKD tell us about the crystallography of the martensitic B2-B19′ transformation in NiTi shape memory alloys. *Crystals* 2020;10:562. DOI
15. Rauch EF, Portillo J, Nicolopoulos S, Bultreys D, Rouvimov S, Moeck P. Automated nanocrystal orientation and phase mapping in the transmission electron microscope on the basis of precession electron diffraction. *Z Kristall* 2010;225:103-9. DOI
16. Rottmann PF, Hemker KJ. Experimental observations of twin formation during thermal annealing of nanocrystalline copper films using orientation mapping. *Scr Mater* 2017;141:76-9. DOI

17. Ma X, Zhao D, Yadav S, Sagapuram D, Xie KY. Grain-subdivision-dominated microstructure evolution in shear bands at high rates. *Mate Res Lett* 2020;8:328-34. DOI

18. Yadav D, Zhao D, Baldwin JK, Devaraj A, Demkowicz MJ, Xie KY. Persistence of crystal orientations across sub-micron-scale "super-grains" in self-organized Cu-W nanocomposites. *Scr Mater* 2021;194:113677. DOI

19. Johnstone DN, van Helvoort ATJ, Midgley PA. Nanoscale strain tomography by scanning precession electron diffraction. *Microsc Microanal* 2017;23:1710-1. DOI

20. Ånes HW, Andersen IM, van Helvoort ATJ. Crystal phase mapping by scanning precession electron diffraction and machine learning decomposition. *Microsc Microanal* 2018;24:586-7. DOI

21. Martineau BH, Johnstone DN, van Helvoort ATJ, Midgley PA, Eggeman AS. Unsupervised machine learning applied to scanning precession electron diffraction data. *Adv Struct Chem Imaging* 2019;3:5. DOI

22. Bergh T, Johnstone DN, Crout P, et al. Nanocrystal segmentation in scanning precession electron diffraction data. *J Microsc* 2020;279:158-67. DOI

23. Portillo J, Rauch EF, Nicolopoulos S, Gemmi M, Bultreys D. Precession electron diffraction assisted orientation mapping in the transmission electron microscope. In: Materials Science Forum; 2010. pp. 1-7. DOI

24. Zhang Y, Fincher CD, Gurrola RM, et al. Strategic texturation of VO_2 thin films for tuning mechanical, structural, and electronic couplings during metal-insulator transitions. *Acta Mater* 2023;242:118478. DOI

25. Zhao D, Patel A, Barbosa A, et al. A reference-area-free strain mapping method using precession electron diffraction data. *Ultramicroscopy* 2023;247:113700. DOI

26. Jani JM, Leary M, Subic A, Gibson MA. A review of shape memory alloy research, applications and opportunities. *Mater Des* 2014;56:1078-113. DOI

27. Schofield P, Bradicich A, Gurrola RM, et al. Harnessing the metal-insulator transition of VO_2 in neuromorphic computing. *Adv Mater* 2022:e2205294. DOI

28. Corti E, Gotsmann B, Moselund K, Ionescu AM, Robertson J, Karg S. Scaled resistively-coupled VO_2 oscillators for neuromorphic computing. *Solid State Electron* 2020;168:107729. DOI

29. de la Pena F, Ostasevicius T, Tonaas Fauske V, et al. Electron microscopy (big and small) data analysis with the open source software package hyperspy. *Microsc Microanal* 2017;23:214-5. DOI

30. Wang L, Zhang Y, Feng J. On the Euclidean distance of images. *IEEE Trans Pattern Anal Mach Intell* 2005;27:1334-9. DOI

31. Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. *Inf Sci* 2015;307:39-52. DOI

32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Pattern Anal Mach Intell* 2004;13:600-12. DOI

33. Lloyd S. Least squares quantization in PCM. *IEEE Trans Pattern Anal Mach Intell* 1982:28;129-37. DOI

34. Gulli A, Pal S. Deep learning with Keras. Birmingham: Packt; 2017.

35. Mika S, Schölkopf B, Smola A, Müller K-R, Scholz M, Rätsch G. Kernel PCA and de-noising in feature spaces. In: advances in neural information processing systems. Cambridge: MIT Press;1999: pp. 536-42.

36. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Stanford, 2006. Available from: https://www.researchgate.net/publication/220778887_K-Means_The_Advantages_of_Careful_Seeding [Last accessed on 27 Jun 2023].

37. Zhao Y, Hwan Lee J, Zhu Y, et al. Structural, electrical, and terahertz transmission properties of VO_2 thin films grown on c-, r-, and m-plane sapphire substrates. *J Appl Phys* 2012;111:053533. DOI

38. Fan LL, Wu YF, Si C, Pan GQ, Zou CW, Wu ZY. Synchrotron radiation study of VO_2 crystal film epitaxial growth on sapphire substrate with intrinsic multi-domains. *Appl Phys Lett* 2013;102:011604. DOI

39. Gonzalez RC, Woods RE. Digital image processing. Pearson education: India; 2009.

40. Dhanachandra N, Manglem K, Chanu YJ. Image segmentation using *k*-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput Sci* 2015;54:764-71. DOI

41. Ray S, Turi RH. Determination of number of clusters in K-means clustering and application in colour image segmentation. In: 4th International Conference on Advances in Pattern Recognition and Digital Techniques. 2023 Jun 17-18; Sydney, Australia. pp. 137-43.

42. Ghahramani Z. Unsupervised learning.In: Advanced Lectures on Machine Learning. 2003 Feb 2-14; ML Summer Schools, Canberra, Australia. pp. 72-112 DOI

43. Ophus C. Four-dimensional scanning transmission electron microscopy (4D-STEM): from scanning nanodiffraction to ptychography and beyond. *Microsc Microanal* 2019;25:563-82. DOI PubMed

44. Williams DB, Carter CB. Transmission Electron Microscopy: Diffraction, Imaging, and Spectrometry. Springer; 2009.

45. Oleynikov P, Hovmöller S, Zou XD. Precession electron diffraction: observed and calculated intensities. *Ultramicroscopy* 2007;107:523-33. DOI PubMed