Clustering Characteristic Diffraction Vectors in 4-D STEM Data Sets from Overlapping Structures in Nanocrystalline and Amorphous Materials

Carter Francis[a], Paul M. Voyles [a]*

[a] Department of Materials Science and Engineering, University of Wisconsin Madison, Madison, Wisconsin 53706, USA

* paul.voyles@wisc.edu

Abstract We describe a method for identifying and clustering diffraction vectors in four-dimensional (4-D) scanning transmission electron microscopy data to determine characteristic diffraction patterns from overlapping structures in projection. First, the data is convolved with a 4-D kernel, then diffraction vectors are identified and clustered using both density-based clustering and a metric that emphasizes rotational symmetries. The method works well for both crystalline and amorphous samples and in high- and low-dose experiments. A simulated dataset of overlapping aluminum nanocrystals provides performance metrics as a function of Poisson noise and the number of overlapping structures. Experimental data from an aluminum nanocrystal sample shows similar performance. For an amorphous $Pd_{77.5}Cu_6Si_{16.5}$ thin film, experiments measuring glassy structure show strong evidence of 4- and 6-fold symmetry structures. A significant background arises from the diffraction of overlapping structures. Quantifying this background helps to separate contributions from single, rotationally symmetric structures vs. apparent symmetries arising from overlapping structures in projection.

Keywords: 4-D STEM; Amorphous Structure; Nanocrystal Characterization; Glass Structure Symmetry; Clustering; Electron Diffraction

1. Introduction

Four-dimensional scanning transmission electron microscopy (4-D STEM) characterizes the structure of materials at nanometer scale and below, providing unique insights into complex materials [1]. In 4-D STEM, a small electron probe rasters across a sample, and a pixelated detector acquires a diffraction pattern in $\boldsymbol{k}$ (k_x, k_y) at each real space position $\boldsymbol{r}$ (r_x, r_y), resulting in a 4-D hypercube $I(r_x, r_y, k_x, k_y)$. Each diffraction pattern $I(k_x, k_y)$ contains information about the structure of the excited volume at position $\boldsymbol{r}$. 4-D STEM experiments characterize local strain and defects [2], crystal orientation [3], polymer orientation [4], and the structure of disordered materials [5–7], among other applications.

Identifying diffraction disks in each diffraction pattern converts the 4-D hypercube into a 4-D set of vectors $<r_x, r_y, k_x, k_y>$, reducing the size and complexity of the data. Methods for disk detection include correlation-based methods [8], circle finding methods, such as variations on the circular Hough transform [9], and feature finding methods such as difference of Gaussians [10]. Convolution neural networks also provide good performance and accuracy for strain and orientation mapping [11]. For thin samples with uniform, well separated disks, any of these methods perform well. Neural networks show good performance in finding diffraction features if the data being processed matches the training dataset [11,12]. In general, methods that depend on edge detection are slightly more accurate for ideal samples while correlation-based methods and

feature finding methods work better for non-ideal samples with varying intensity in the diffracting disks.

Strain mapping and vector-based orientation mapping are straightforward calculations given a list of vectors. The relative shifts in the diffraction vector relate to local lattice distortions along different directions in the material [2]. Orientation maps of crystalline materials rely on the comparison of the list of diffraction vectors with the list of vectors from a library of values [13]. In comparison to image cross correlation methods [14], vector-based methods offer higher speed and more flexibility at the cost of additional pre-processing, specifically finding diffraction vectors, which leads to less automation.

More complex analyses calculate characteristic diffraction patterns from structures probed in a 4-D STEM dataset. Methods include matrix factorization [15], template matching versus a library of pre-computed structures using traditional methods or neural networks [16], and density-based clustering [17]. Matrix factorization segments data quickly without preprocessing. It excels at identifying components but struggles to separate structures that overlap through the thickness of the sample and tends to over-cluster in many situations [15,18]. Template matching proves very effective for 1-3 overlapping structures through the thickness [13]. Knowledge about the underlying crystal structure improves clustering performance but limits the ability to measure unknown structures [12]. Vector based methods require less memory and are less computationally expensive than methods that use every pixel in the 4-D hypercube, but reliably finding diffraction vectors in an automated fashion can be difficult [10]. In every method, clustering performance decreases significantly as the number of overlapping structures increases.

4-D STEM is widely used for measuring the structure of amorphous materials, but data from current methods for measuring amorphous structure can be difficult to interpret. These methods include fluctuation electron microscopy (FEM), angular correlations/angular power spectrum (AC/APS), and correlation symmetry analysis (CSA). FEM measures the global variation in the structure as a function of $\boldsymbol{k}$ [5,19]. Methods for correcting for thickness extend the functionality of FEM to samples of up to 1.5 times the elastic mean free path [20]. AC or correlographs provide additional information about the symmetry of the diffracting structures, but overlapping structures complicate the connection between observed symmetries in diffraction and symmetries in atomic structure [6,21,22]. CSA is, in many cases, superior to AC with respect to overlapping structures, but as samples get thicker, identifying structural symmetry and local structure with high confidence is still difficult [7].

Here, we present a method for identifying diffraction disks in 4-D STEM data that is more robust than previous methods against overlapping structures and shows good performance against Poisson noise. The method first filters the intensity in four dimensions, then finds diffraction vectors in each diffraction pattern to generate the list of vectors $<r_x, r_y, k_x, k_y>$. Density-based clustering on the vector list is used to determine the characteristic diffraction patterns for individual structures, separating out structures that partially but not completely overlap in real space. This method performs well for crystalline and amorphous materials and provides a general workflow for important feature identification in 4D STEM datasets regardless of the sample structure. Tests on phantom data of Al nanocrystals show good performance versus Poisson noise and the number of overlapping structures. Experiments on a real Al nanocrystals sample show similarly good performance. For a metallic glass sample, adjusting the intensity threshold for diffraction vector identification allows us to identify speckles from high symmetry structures with high confidence.

Measuring the angle between speckles at fixed radius in the diffraction pattern, k, provides insight into the symmetry of the dominant structures in the glass.

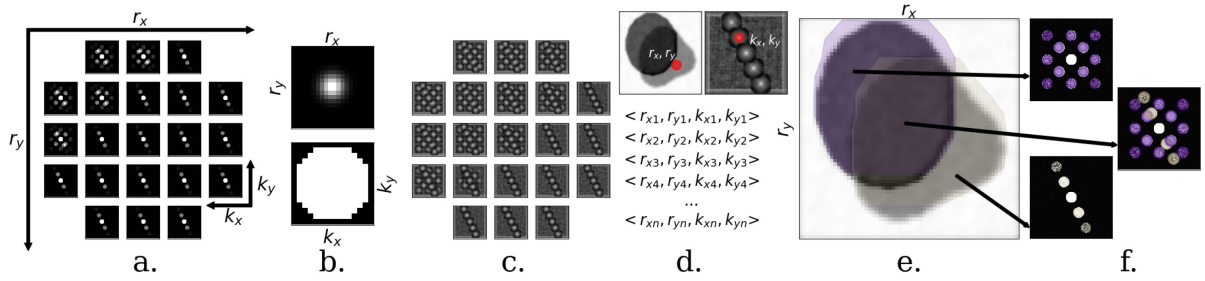

Figure 1: Graphical representation of the filtering - peak finding - segmentation workflow for a simulated dataset from a simple sample consisting of two overlapping Al nano-crystals. (a) example diffraction patterns. (b) convolution kernel in real and reciprocal space. (c) data from (a) after filtering. (d) diffraction vectors identified from (c). (e) Overlapping but distinct structures identified in real space. (f) diffraction patterns from the indicated positions, with each feature color coded by its corresponding real space object.

2. Methods

Figure 1 illustrates the vector-finding and clustering method for a sample example consisting of two partially overlapping Al nanocrystals. Figure 1a shows diffraction patterns from the overlapping region of two crystals. The upper left of the tableau shows the superposition of the two characteristic diffraction patterns, while the lower left shows only one characteristic diffraction pattern.

Figure 1b shows the kernel that is used to filter the data using the normalized cross correlation. The kernel is Gaussian in $\boldsymbol{r}$ and a flat disk in $\boldsymbol{k}$ to match the probe intensity in both spaces. The size of the kernel in $\boldsymbol{r}$ and $\boldsymbol{k}$ is also selected to match the probe intensity. Figure 1c shows the resulting data after filtering. Filtering suppresses features in real space smaller than the real-space size of the beam [23] and accentuates features in reciprocal space with a similar diameter to the probe in reciprocal space. After filtering, we apply a simple local maximum peak finding method. A threshold of around 50% of the maximum cross-correlation intensity returns the positions of the diffraction vectors shown in Figure 1d. Here, we use the term "diffraction vector" to refer to either the 4-D vectors or, occasionally, the (k_x, k_y) position of an identified feature in a particular diffraction pattern.

We then apply a two-step density-based clustering method to identify real space structures and their corresponding diffraction patterns. First, we use density-based spatial clustering of applications of noise (DBSCAN) to identify dense clusters of 4-D vectors $<r_x, r_y, k_x, k_y>$. To cluster successfully, the real and reciprocal space magnitudes must be scaled independently. An ideal cluster covers all the r_x, r_y positions associated with a given structure, but only small range of k_x, k_y associated with the center of a single diffraction disk. DBSCAN is controlled by the adjustable parameters ϵ, the radius of the search for some hypersphere, and m, the minimum number of vectors in the hypersphere to identify a cluster. Clusters are agglomerated from overlapping hyperspheres. m is typically set to the number of dimensions + 1, but since our k_x, k_y clusters should

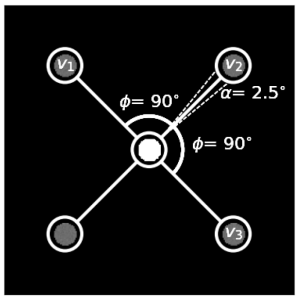

Figure 2: Schematic of the 3-vector clustering applied to complex overlapping structures. Only sets of 3 vectors subtending two equal angles, ϕ, within an error of $\pm\alpha$ are considered

be small, we set m = 2+1. We then scale the 4-D vectors such that ϵ is ~1 pixel in k_x, k_y corresponding to a small uncertainty in the position of the center of the identified diffraction disks, and, at the same time, ~2 pixels in r_x, r_y to account for potential a diffraction disk to missed due to noise or other factors from a pattern that nonetheless arises from the same real space object as its neighbors.

Second, we compute for each cluster the average and standard deviation for r_x, r_y and (separately) for k_x, k_y. Clusters with a large standard deviation in k_x, k_y are flagged for further analysis. Third, we use DBSCAN again to cluster only the r_x, r_y average vectors. These average vector clusters represent crystals with the same orientation. For each crystal, we construct a list of all the k_x, k_y average vectors from the associated 4-D clusters. This list makes up the characteristic diffraction pattern from some crystal. Figures 1e and 1f show the results of clustering the example data. Figure 1e shows the real space regions of the two nanocrystals in the dataset, and Figure 1f shows the characteristic diffraction patterns color-coded by which disks arise from each crystal, even in the region where the two crystals overlap.

Figure 2 illustrates the symmetry-based clustering metric used to identify high-symmetry prototypical diffraction patterns. Symmetric diffraction patterns are defined by having at least 3 vectors at the same real space position [$<r_x, r_y, k_{x1}, k_{y1}>$, $< r_x, r_y, k_{x2}, k_{y2}>$, $< r_x, r_y, k_{x3}, k_{y3}>$] that at least twice subtend the angle ϕ, within an acceptance angle α. *If* $\alpha > |\phi_1 - \phi_2|$, then ϕ_1 is considered equal to ϕ_2. For this study an $\alpha = 2.5°$ was used. The angle ϕ is used to describe the symmetry of the 3 vectors. DBSCAN can be used to cluster based on the computed $< r_x, r_y, \phi>$ vector to identify a characteristic diffraction pattern from a highly symmetric structure.

This method was tested on phantom simulated data and experimental data. The phantom data was simulated from models of randomly oriented Al nanocrystals using kinematic diffraction simulations implemented in the diffsims python package. A maximum excitation error of 0.015 $Å^{-1}$ and beam voltage of 200 keV were used. Each diffraction vector was represented by a flat disk with a radius of 5 pixels or 0.05 $Å^{-1}$. Poisson and Gaussian noise were added to simulate the electron dose and detector readout noise respectively. The use of phantom data with known ground truth enables us to calculate recall percentages as a measure of the method performance versus the number of structures and Poisson noise. We define two different recall percentages: Diffraction vector recall measures the percent of diffraction vectors correctly identified. A diffraction vector is recalled if the algorithm identifies a diffraction vector within 1 pixel of the ground truth.

Diffraction pattern recall measures the percent of characteristic diffraction patterns correctly identified. Because only some of the diffraction vectors of some characteristic diffraction pattern may be identified, for instance, due to noise, both partial and complete pattern recall are reported.

A FEI Titan electron microscope was used to acquire a 4-D STEM dataset from an evaporated Al nanocrystal sample (Electron Microscopy Supplies, part #80044). The experiment was done at spot size 5 and with a convergence angle of 1.5 mrad which created a probe with a diameter of around 1 nm. 1024x1024 diffraction patterns were acquired over a 522 nm x 522 nm area, giving a step size of 0.51 nm. The data was acquired using a Direct Electron Celeritas camera with a readout speed of 20,000 frames per second over a period of around 1 minute to minimize drift. The 4-D filter used for this data had a real-space width $\sigma = 1.0$ nm and a reciprocal space disk size of 1.5 mrad.

The Titan STEM was also used to acquire 4-D STEM data on $Pd_{77.5}Cu_6Si_{16.5}$ thin film glasses of various thicknesses at spot size 5, a convergence angle of 2.5 mrad, and probe size of 0.5 nm, rastered across the sample with a 0.1 nm step size. The real space sampling was 1024x1024 diffraction patterns and the data was acquired using a Direct Electron Celeritas camera with a readout speed of 5,000 frames per second. The resulting datasets were filtered, and peaks within the dataset were found. Then, high symmetry diffraction patterns were identified using the symmetry clustering method described above. These clusters were then further clustered into structures diffracting from some high symmetry axis using DBSCAN. Processing was done using the hyperspy[24], pyxem[25] and diffsims[26] packages.

Processing was done lazily (from hard disk) using distributed computing resources (128 Cores with 512 GB RAM). The total time to process the large (256 GB) Al dataset was around ~7 minutes. Processing was also tested and ran effectively when running lazily on a MacBook Air M1 laptop with 16 GB of RAM and 8 cores.

3. Results

Figure 3 shows the diffraction vector recall percentage as a function of the number of crystals in the phantom nanocrystalline Al 4-D STEM data set with minimal noise. The recall percentage falls as the density of nanocrystals increases. As the number of nanocrystals in projection increases, more diffraction disks overlap by chance and are not identified in the diffraction pattern. For the largest number of crystals tested here (1200 in the field of view), the average number of structures in projection is ~7, and the vector recall percentage falls to 83%. When only considering the strongest 50% of diffracting vectors at an average of ~7 structures in projection, the vector recall percentage remains high, with ~95% of the vectors successfully identified.

Figure 4 shows the diffraction vector recall percentage as a function of Poisson noise, controlled by the average number of electrons per diffraction vector. The synthetic data has 40 nanocrystals in the field of view. The recall percentage is above 90% for 13 or more electrons per diffraction disk but falls quickly at lower numbers of electrons. The top axis displays the total number of electrons within the 4-D volume covered by the kernel used for filtering. As the Gaussian blur contains information from surrounding diffraction patterns in (r_x, r_y) and the normalized cross correlation uses the entire diffraction disk, this effective number of electrons better represents the total electron dose used to identify the diffraction vector.

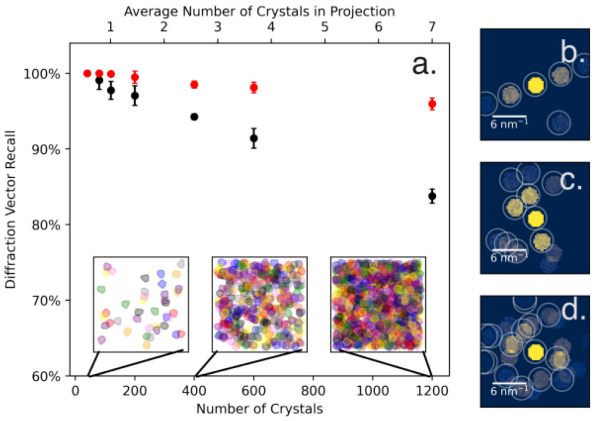

Figure 3: (a) The percentage of diffraction vectors recalled using the filtering and peak finding method as a function of the number of crystals in the field of view. The inset figures show real space representations of selected nanocrystalline Al models. The red markers show only the strongest diffracting 50% of vectors where the black markers show all vectors regardless of intensity. (b) An example diffraction pattern for an average number of crystals in projection of less than 1. (c) An example diffraction pattern for an average number of crystals in projection of 2.6. (d) An example diffraction pattern for an average number of crystals in projection of 7.

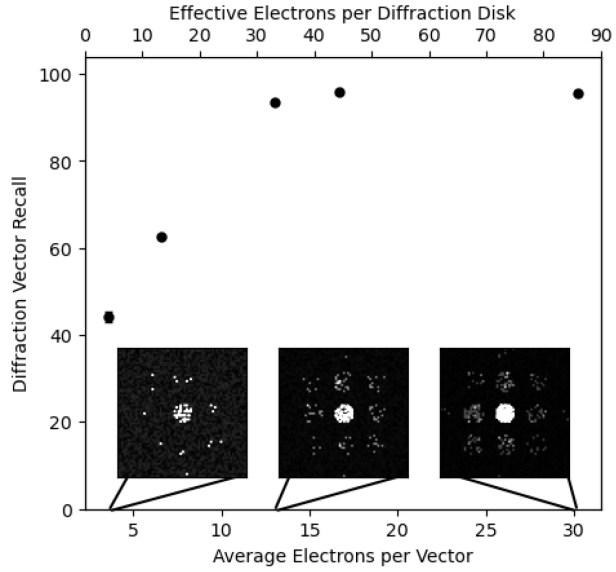

Figure 4: Diffraction vector recall percentage as a function of the average number of electrons per vector for synthetic data with 40 nanocrystals. The top axis shows the effective number of electrons used to find each diffraction vector after 4-D filtering, which combines counts in reciprocal and real space. The inset images are example single diffraction patterns with an average of 2,12, and 30 electrons per diffraction disk.

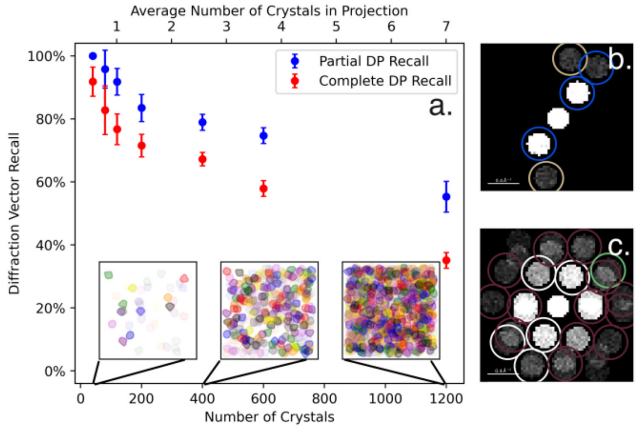

Figure 5: (a) Recall percentage for complete diffraction patterns after clustering with 2-step density-based clustering. Complete diffraction pattern recall refers to every diffraction vector for a crystal being correctly found. Partial diffraction pattern recall refers to only some of the diffraction vectors for a crystal being correctly found**.** (b) Example of complete pattern recall of two crystals, one contributing the spots circled in blue, the other contributing the spots circled in yellow. **(c)** Example of partial pattern recall, with the white circles identifying vectors that are not properly clustered due to overlapping crystals in projection. The magenta and green circles show a partially recalled high symmetry diffraction pattern and a partially recalled off-axis diffraction pattern.

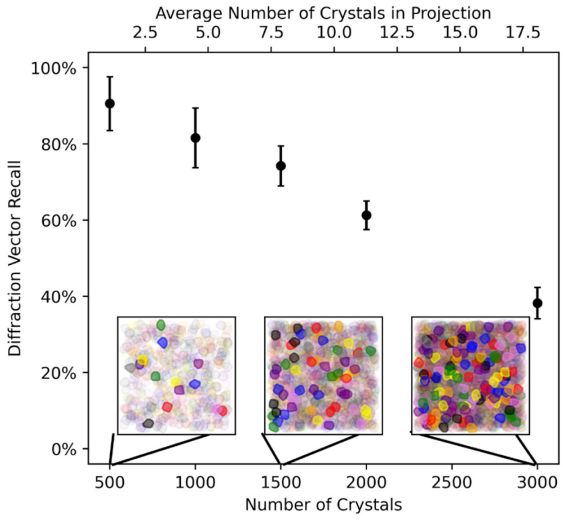

Figure 6: The diffraction pattern recall percentage using the three-vector separation clustering method. Recall is only measured for the crystals oriented such that they diffract on a high symmetry axis. The inset images show a real space image of all the crystals in the dataset, with

Figure 5a shows the diffraction pattern recall for characteristic diffraction patterns with two or more diffraction vectors. Complete diffraction pattern recall measures the number of characteristic diffraction patterns for which every diffraction vector is correctly identified and clustered. Partial diffraction pattern recall measures the number of characteristic diffraction patterns for which 2 or

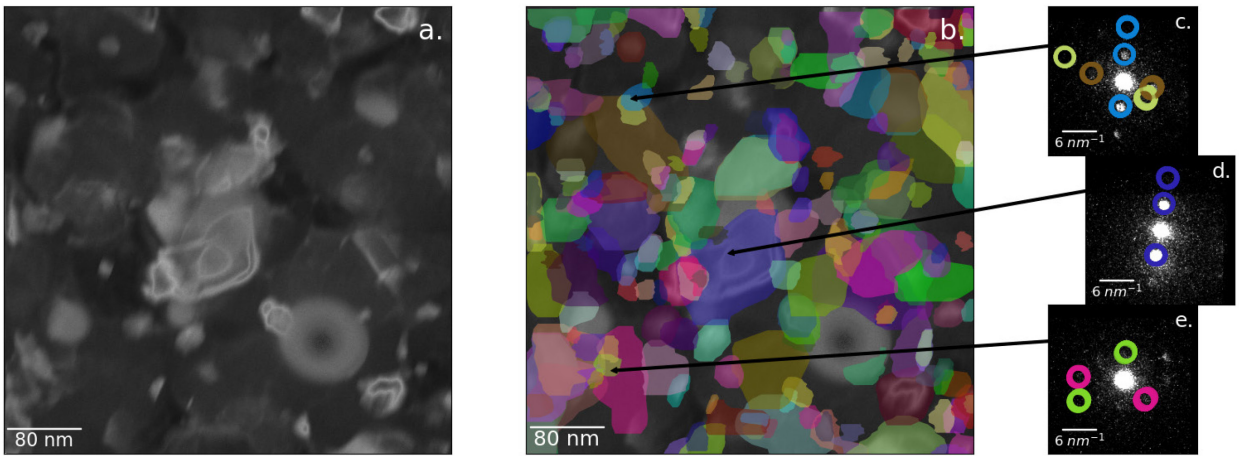

Figure 7: Clustering 4-D STEM experimental data from an Al nanocrystal thin film sample. (a) A low angle virtual dark field image of the first two diffraction rings (b) the segmented nanocrystals. (c)-(e) diffraction patterns from positions on the sample and the resulting clustering. For (a) the ring in the lower right is a result of carbon contamination. It does not interfere with the crystal identification. Diffraction vectors in c and e which are not circled are identified but not correctly clustered.

more diffraction vectors are identified and clustered but less than the number of diffraction vectors in the complete diffraction pattern. For up to 600 crystals in the field of view, the partial recall remains >80%. For orientation mapping or measuring structure symmetry, higher pattern recall will lead to better performance. Figure 5b shows an example of complete pattern recall and Figure 5c shows an example of partial pattern recall with the white circles showing vectors identified but not properly clustered due to overlapping nanocrystals. The colored circles show diffraction patterns from individual overlapping crystals. In Figure 5b, two separate crystals are completely recalled, while in Figure 5c shows partial recall due to high symmetry diffraction with many overlapping disks and structures. Interactive visualization of this phenomenon is available with the included Jupyter notebooks.

Figure 6 shows the diffraction pattern recall after filtering based on the three vector separation criteria described above. The ground truth for this pattern recall percentage only includes patterns with three or more vectors separated by an angle ϕ. The low probability of three randomly oriented vectors potentially arising from two different but overlapping crystals, satisfying this condition significantly improves the diffraction pattern recall performance. The inset figures show crystals that meet this condition in bold overlaid on the total number of crystals. The smaller number of crystals involved in the analysis reduces random overlaps in space and increases the pattern recall percentage. In Figure 6 the recall percentage at 1200 crystals remains around 80% rather than 40% as in Figure 5. The recall percentage in Figure 6 does not fall to 40% until the sample consists of 3000 crystals.

Figure 7 shows diffraction vector identification and clustering on experimental 4-D STEM data acquired from the Al nanocrystal sample. Figure 7a is a virtual dark-field image computed from the 4-D STEM data, the circular object in the bottom right corner arises from sample contamination. Figure 7b shows the clustered and segmented crystals. Figures 7c-e show example diffraction patterns from different positions on the sample. In each pattern, diffraction spots are

color coded by the crystal from which they arise. The average diameter of the nanocrystals is ~36 nm, and, based on the segmentation results, the average number of overlapping crystals through the thickness is 2-4. Based on these parameters, the complete diffraction pattern recall for the dataset should be 70-80%, based on Figure 5. Visual inspection of the diffraction patterns and segmentation compared to images like Figure 7a seems consistent with 70-80% recall. The 4-D STEM data analyzed to produce Figure 7 is 256 GB of raw images. Identifying diffraction vectors drastically reduces the data size, to ~150 MB.

We have also compared the three-vector direct symmetry analysis with previous methods for measuring symmetry in amorphous materials. Figure 8a shows the true positive recall for identifying six-fold symmetric structures within the toy dataset as a function of the number of structures within the field of view. Figure 8b shows the false positive recall for the same dataset. The direct symmetry analysis presented here indicates higher performance concerning overlapping structures than the correlation symmetry analysis and the angular correlation/ power spectrum analysis. Similar results are seen for 4-fold symmetry as well.

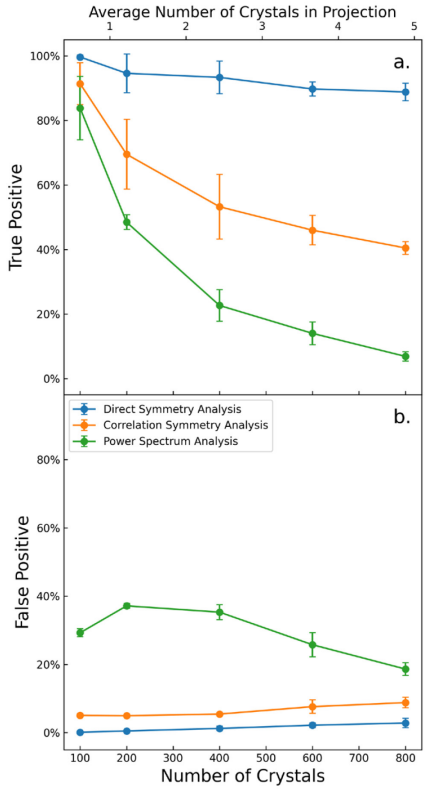

Figure 8: Comparison of recall performance for spatial structure identification for 6-fold symmetric structures in the toy Al dataset. (a) The true positive recall. (b) The false positive recall. Power spectrum analysis, correlation symmetry analysis, and direct symmetry analysis (described here) are all compared. Thresholds are chosen to maximize true positive recall for the power spectrum and the correlation symmetry analysis.

Figure 9 shows the dominant symmetries in the PdCuSi glass sample found using the 3-vector method in the form of histograms of the number of diffraction patterns exhibiting sets of three or more vectors subtending different angles, ϕ. Different histograms were computed using different thresholds in the peak finding step. Strong 12-, 6- and 4-fold symmetries in the diffraction patterns give rise to the peaks in the histograms at $\pi/6$, $\pi/3$ and $\pi/2$. The relative fraction of 6-fold structures is higher with a higher threshold, and the relative fractions of 12- and 4-fold structures are lower. As the threshold increases the number of randomly overlapping structures decreases. Figure 9 was produced from one 64 GB 4D STEM dataset. Identifying diffraction vectors drastically reduces the data size, to ~100 MB for the lowest peak identification threshold in 9a to ~10 MB for the highest peak identification threshold in Figure 9(e).

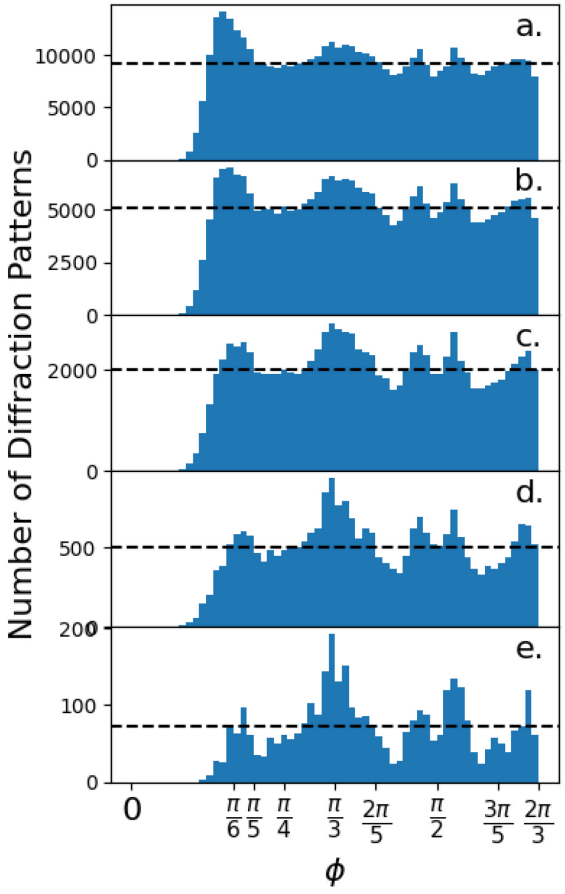

Figure 9: Histograms of the number of diffraction patterns as a function of the angular separation ϕ for the 3-vector clustering method for the 5 nm thick PdCuSi glass sample. (a) histogram for a low threshold for peak finding. (b)-(e) histograms for progressively higher thresholds. The dotted lines represent the symmetry expected based on random overlaps.

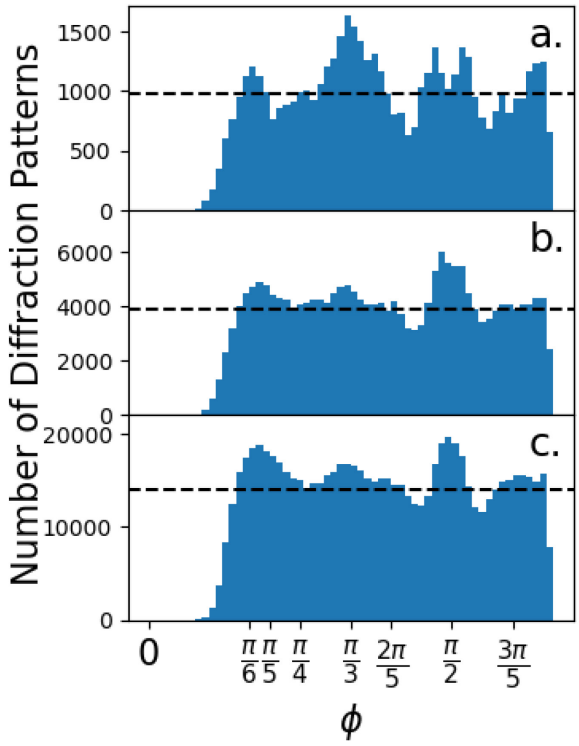

Figure 10: Histograms of the number of diffraction patterns as a function of the separation ϕ for the 3-vector clustering method. (a) histogram for a 5 nm thickness PdCuSi glass (b) 9 nm thickness, and (c) 13 nm thickness. The dotted lines represent the symmetry expected based on random overlaps.

Figure 10 is a histogram of the number of symmetric diffraction patterns as a function of subtended angle identified in the PdCuSi glass sample as a function of the sample thickness from 5 to 13 nm. Increasing the sample thickness should increase the number of overlapping diffracting structures contributing to each experimental diffraction pattern. Increasing thickness changes the histograms similarly to decreasing the peak finding threshold. The relative fraction of 6-fold structures increases with a lower thickness, and the relative fraction of 12- and 4-fold structures decreases as the thickness increases.

4. Discussion

Previous results have shown that using a known kernel, such as a vacuum probe, is an excellent method for extracting the position of diffraction vectors [8]. These works primarily focus on orientation mapping and strain mapping with simple, non-overlapping, structures in real space projection [11,13]. We found convolution with a known kernel remains effective for samples with overlapping structures in real space and overlapping disks in reciprocal space. Figure 3 shows that the recall percentage for the diffraction vectors remains high even with larger numbers of diffracting structures with a significant number of through-thickness overlaps at each real space position. The window-normalized cross-correlation specifically proves the most effective, as it normalizes based on the local intensity. This helps to identify both low and high-intensity diffraction vectors. One potential breakdown of this method occurs when diffraction spots significantly overlap in reciprocal space and have large differences in intensity. In this case, the normalization suppresses the weaker diffracting vector. When the method fails, lower-intensity diffraction vectors are not properly identified. This can be seen by the difference in the red and black markers in Figure 3. Overlap in reciprocal space can, of course, be minimized using a small probe convergence angle at the price of a larger probe and poorer spatial resolution.

Figure 4 shows this method is robust against noise. Compared to the position error published in Pekin et. al. [8], this method performs optimally at 1/10th of the electron dose per diffraction pattern. The robust performance results from the Gaussian filter applied in r to the dataset, which reduces Poisson noise but reduces the real space resolution. For most mapping measurements, increasing the number of counts in the dataset is better than applying the Gaussian filter, if the sample will sustain the dose. For low-dose strain mapping, however, even a small sigma value ($\sigma = 0.25$ nm) increases the number of effective counts and reduces readout noise while only slightly reducing the spatial resolution. This is equivalent to increasing the size of the electron beam to $\sqrt{\sigma_{\text{beam}}^2 + \sigma_{\text{filter}}^2}$ with the added benefit of suppressing the readout noise and reducing the overall dose required for the experiment/applied to the sample.

Figure 5 shows additional difficulties that arise from overlapping structures in real space and in reciprocal space. In both cases, the greedy DBSCAN method tends to over-cluster, leading to poorer results as overlaps increase. Separating the diffraction patterns from structures that completely overlap in real space requires some additional prior information about the material being measured. We explored using rotational symmetry to cluster diffraction vectors, but future work using both density-based clustering and a library of crystal orientations might prove useful for clustering overlapping crystals. Figure 6 shows a marked increase in recall achieved by focusing only on higher symmetry diffraction patterns. Emphasizing symmetric features reduces

the over-clustering since it is relatively rare for three diffraction vectors to subtend the same angle ϕ and for all three vectors to extend for the same region of real space.

Many of the characteristic diffraction patterns for the experimental Al nanocrystal sample in Figure 7 show physically reasonable features, including strong Friedel symmetry and rotational symmetry. This result suggests that the 2-step clustering does a good job of separating overlapping crystals in the sample, but the greedy nature of the spatial clustering means that small variations in the positions of the diffraction vectors from small tilts or strains are kept within the same crystal. Vector-based template matching based on prior knowledge of the crystal structures in the sample is likely to provide even better separation of overlapping crystals using methods like those discussed by Valery et. al. and Ophus et.al. [13,27]. Such methods are readily applied to the list of 4D vectors extracted from the dataset before 2-step clustering. The 2-step clustering method presented here is particularly useful for samples with unknown crystal structures, samples that contain as-yet unidentified second phases, or amorphous samples.

Overlapping structures in projection have been recognized for some time as a challenge for assessing rotational symmetries in diffraction from amorphous materials. Gibson, Treacy, and Tao suggested that only 2-fold symmetry was robust enough to be reliably measured [28]. Im et. al. showed that thicker samples result in both even and odd symmetries in the AC [22]. The magnitude of the correlation in AC is especially difficult to interpret directly. Corrections from Liu provide some insight into connecting specific glassy structures with angular correlations, but they do not account for overlapping structures [6]. CSA reduces the effects of overlapping structures in comparison to AC by probing a narrow range of angles near the angles defined by a perfect rotational symmetry [7]. Some models for correcting overlaps have been proposed based on the size of the structures [29] but break down with increasing numbers of overlapping structures as shown in Figure 8.

Figure 10 shows that as the number of structures through the thickness increases, symmetries from random overlaps increase and a flat, ϕ-independent background arises. Figure 9 shows that increasing the threshold for peak finding reduces random overlaps but maintains strong symmetry recovery in complex disordered materials. Unlike previous AC or CSA methods, the probability for random symmetries from overlapping structures in the current method can be calculated. For any set of 2 vectors there is a $2\alpha/\pi$ chance that a randomly oriented third vector will satisfy the 3-vector condition, where α is the acceptance angle. This effect gives rise to an average number of false positive structures given by $C(n,3) * 2\alpha/\pi$ where $C(n,3)$ is the number of combinations for n vectors in each diffraction pattern in groups of 3. The dotted lines in each panel are calculated from this formula, showing that a large fraction of the total symmetric patterns arise from random overlaps when the number of diffraction vectors is high. Thin samples and high thresholds reduce this effect at the cost of ignoring weakly diffracting structures. If the angular symmetries were assessed from just two diffraction vectors instead of three, the background level would be significantly higher at $C(n,2)$. This explains why this method outperforms correlation symmetry analysis and angular power spectrum methods as shown in Figure 8. Both methods only depend on a two-angle comparison. Additionally, the performance of the power spectrum method is further negatively impacted by the Fourier ringing artifact described by Huang [7]. As a result, all structures with Friedel symmetry show strong 2*n fold symmetry, and the number of false positives is much higher than other methods.

For thick samples with a characteristic symmetry at ϕ, a secondary peak at $\phi/2$ arises with a frequency $P(\phi) * n * 2\alpha/\pi$ where n is the number of vectors and $P(\phi)$ is the probability that a group of 2 vectors has symmetry ϕ. This effect explains the 12-fold symmetry in Figures 8 and 9. The relative height of the 12-fold peak with respect to the 6-fold peak increases with increasing thickness and with decreasing threshold, both of which increase n. Similarly, the strong 4-fold symmetry in the glass might partially arise from the strong 2-fold Friedel symmetry in the diffraction patterns. This false positive contribution is not shown on the figures as it requires knowledge of the true, ground truth $P(\phi)$, which is not available for experimental data.

CSA measured for this same glass by Huang et. al. shows the presence of both 4- and 6-fold symmetries [29]. Here we show similar structures but suggest that some of the strong 4-fold symmetries arise from random overlaps with strong Friedel pairs. Additionally, this study shows little evidence of 10-fold symmetries measured with CSA. The current analysis has the advantage of identifying the specific diffraction vectors and characteristics that contribute to specific symmetries, which may open the door to additional analysis of selected, highly symmetrical diffraction patterns.

The small size of the list of 4D vectors compared to the original data makes it amenable to additional processing beyond what is discussed here. For example, double diffraction from partially overlapping structures in crystalline samples may lead to incorrect identification of separate spatial regions with diffraction vectors that are a sum of a vector from the top crystal and a vector from the bottom crystal. Further processing might identify such regions by finding diffraction vectors are the sum of a vector from one spatial region and a vector from a different spatial region. These regions might also be identified by a failure to match against a simulated library as in orientation mapping. Other dynamical diffraction phenomena, like the appearance of kinematically forbidden reflections, similarly will result in 4D vectors that require additional processing based on prior knowledge of the sample to identify.

5. Conclusions

We have developed a method for analyzing 4-D STEM data that consists of filtering, peak finding to identify 4-D diffraction vectors, then clustering those vectors into characteristic diffraction patterns. This approach is well suited to characterizing diffraction from spatially overlapping structures along the electron beam, such as nanocrystalline materials and amorphous materials. Density based clustering is adequate for extracting characteristic diffraction patterns of individual crystals in data acquired from nanocrystalline material. Clustering of more complex glassy materials benefits from symmetry-based clustering based on 3 vector criteria. The method identifies structural information from glassy samples of 5-13 nm in thickness. Good performance on low dose or sparse data allows for fast acquisition of large data sets, which could potentially cover large areas or arise from time resolved experiments. The method results in a large reduction in data size, which enables more facile complex analysis.

6. Data and Software Availability

The methods for the analysis are all available in the HyperSpy [10.5281/zenodo.10412190] and pyxem[10.5281/zenodo.10551678] software packages.

Methods for simulating the toy models are also available in the diffsims [10.5281/zenodo.7962969] software package.

The 4-D STEM data analyzed are available hosted from the Materials Data Facility[10.18126/z6m9-o3hv]

Jupyter Notebooks for running the analysis are hosted at:

https://github.com/CSSFrancis/4d_stem_clustering

7. Acknowledgements

This research was supported by the National Science Foundation (DMR-2204632). Experimental 4-D STEM data were collected using facilities support by the Wisconsin MRSEC (DMR-230900).

8. References

[1] C. Ophus, Four-Dimensional Scanning Transmission Electron Microscopy (4D-STEM): From Scanning Nanodiffraction to Ptychography and Beyond, Microscopy and Microanalysis (2019). https://doi.org/10.1017/S1431927619000497.

[2] A. Armigliato, S. Frabboni, G.C. Gazzadi, Electron diffraction with ten nanometer beam size for strain analysis of nanodevices, Appl Phys Lett 93 (2008). https://doi.org/10.1063/1.3003581.

[3] E.F. Rauch, J. Portillo, S. Nicolopoulos, D. Bultreys, S. Rouvimov, P. Moeck, Automated nanocrystal orientation and phase mapping in the transmission electron microscope on the basis of precession electron diffraction, Zeitschrift Fur Kristallographie 225 (2010) 103–109. https://doi.org/10.1524/zkri.2010.1205.

[4] K.C. Bustillo, O. Panova, X.C. Chen, C.J. Takacs, J. Ciston, C. Ophus, N.P. Balsara, A.M. Minor, Nanobeam Scanning Diffraction for Orientation Mapping of Polymers, Microscopy and Microanalysis 23 (2017) 1782–1783. https://doi.org/10.1017/s1431927617009576.

[5] P.M. Voyles, D.A. Muller, Fluctuation microscopy in the STEM, Ultramicroscopy 93 (2002) 147–159.

[6] A.C.Y. Liu, G.R. Lumpkin, T.C. Petersen, J. Etheridge, L. Bourgeois, Interpretation of angular symmetries in electron nanodiffraction patterns from thin amorphous specimens, Acta Crystallogr A Found Adv 71 (2015) 473–482. https://doi.org/10.1107/S2053273315011845.

[7] S. Huang, C. Francis, J. Ketkaew, J. Schroers, P.M. Voyles, Correlation symmetry analysis of electron nanodiffraction from amorphous materials, Ultramicroscopy 232 (2022) 113405. https://doi.org/10.1016/j.ultramic.2021.113405.

[8] T.C. Pekin, C. Gammer, J. Ciston, A.M. Minor, C. Ophus, Optimizing disk registration algorithms for nanobeam electron diffraction strain mapping, Ultramicroscopy 176 (2017) 170–176. https://doi.org/10.1016/j.ultramic.2016.12.021.

[9] L. Wu, M.G. Han, Y. Zhu, Toward accurate measurement of electromagnetic field by retrieving and refining the center position of non-uniform diffraction disks in Lorentz 4D-STEM, Ultramicroscopy 250 (2023). https://doi.org/10.1016/j.ultramic.2023.113745.

[10] D.N. Johnstone, B.H. Martineau, P. Crout, P.A. Midgley, A.S. Eggeman, Density-based clustering of crystal (mis)orientations and the orix Python library, J Appl Crystallogr 53 (2020) 1293–1298. https://doi.org/10.1107/S1600576720011103.
[11] R. Yuan, J. Zhang, L. He, J.M. Zuo, Training artificial neural networks for precision orientation and strain mapping using 4D electron diffraction datasets, Ultramicroscopy 231 (2021). https://doi.org/10.1016/j.ultramic.2021.113256.
[12] C. Ophus, A. Rakowski, J. Munshi, B.H. Savitzky, S. Zeltmann, A. Bruefach, M. Scott, J. Ciston, A.M. Minor, M.K. Chan, 4D-STEM Analysis with the Open Source py4DSTEM and crystal4D Toolkits, Microscopy and Microanalysis 28 (2022) 3054–3055. https://doi.org/10.1017/s1431927622011394.
[13] C. Ophus, S.E. Zeltmann, A. Bruefach, A. Rakowski, B.H. Savitzky, A.M. Minor, M.C. Scott, Automated Crystal Orientation Mapping in py4DSTEM using Sparse Correlation Matching, Microscopy and Microanalysis 28 (2022) 390–403. https://doi.org/10.1017/S1431927622000101.
[14] E.F. Rauch, J. Portillo, S. Nicolopoulos, D. Bultreys, S. Rouvimov, P. Moeck, Automated nanocrystal orientation and phase mapping in the transmission electron microscope on the basis of precession electron diffraction, Zeitschrift Für Kristallographie 225 (2010) 103–109. https://doi.org/10.1524/zkri.2010.1205.
[15] A. Bruefach, C. Ophus, M.C. Scott, Robust design of semi-automated clustering models for 4D-STEM datasets, APL Machine Learning 1 (2023) 016106. https://doi.org/10.1063/5.0130546.
[16] E. Thronsen, T. Bergh, T.I. Thorsen, E.F. Christiansen, J. Frafjord, P. Crout, A.T.J. Van Helvoort, P.A. Midgley, R. Holmestad, Scanning precession electron diffraction data analysis approaches for phase mapping of precipitates in aluminium alloys, Ultramicroscopy 255 (2024) 304–3991. https://doi.org/10.5281/zenodo.6.
[17] T. Bergh, D.N. Johnstone, P. Crout, S. Høgås, P.A. Midgley, R. Holmestad, P.E. Vullum, A.T.J.V. Helvoort, Nanocrystal segmentation in scanning precession electron diffraction data, J Microsc 279 (2020) 158–167. https://doi.org/10.1111/jmi.12850.
[18] F.I. Allen, T.C. Pekin, A. Persaud, S.J. Rozeveld, G.F. Meyers, J. Ciston, C. Ophus, A.M. Minor, Fast Grain Mapping with Sub-Nanometer Resolution Using 4D-STEM with Grain Classification by Principal Component Analysis and Non-Negative Matrix Factorization, Microscopy and Microanalysis 27 (2021) 794–803. https://doi.org/10.1017/S1431927621011946.
[19] T.L. Daulton, K.S. Bondi, K.F. Kelton, Nanobeam diffraction fluctuation electron microscopy technique for structural characterization of disordered materials-Application to Al88-xY7Fe5Tix metallic glasses, Ultramicroscopy 110 (2010) 1279–1289. https://doi.org/10.1016/j.ultramic.2010.05.010.
[20] F. Yi, P.M. Voyles, Effect of sample thickness, energy filtering, and probe coherence on fluctuation electron microscopy experiments, Ultramicroscopy 111 (2011) 1375–1380. https://doi.org/10.1016/j.ultramic.2011.05.004.
[21] T. Sun, M.M.J. Treacy, T. Li, N.J. Zaluzec, J.M. Gibson, The importance of averaging to interpret electron correlographs of disordered materials, Microscopy and Microanalysis 20 (2014) 627–634. https://doi.org/10.1017/S1431927613014116.
[22] S. Im, Z. Chen, J.M. Johnson, P. Zhao, G.H. Yoo, E.S. Park, Y. Wang, D.A. Muller, J. Hwang, Direct Determination of Structural Heterogeneity in Metallic Glasses Using Four-

Dimensional Scanning Transmission Electron Microscopy, Ultramicroscopy (2018). https://doi.org/10.1016/j.ultramic.2018.09.005.
[23] H. Kong, H.C. Akakin, S.E. Sarma, A generalized laplacian of gaussian filter for blob detection and its applications, IEEE Trans Cybern 43 (2013) 1719–1733. https://doi.org/10.1109/TSMCB.2012.2228639.
[24] F. de la Peña, E. Prestat, V.T. Fauske, P. Burdet, J. Lähnemann, P. Jokubauskas, T. Furnival, M. Nord, T. Ostasevicius, K.E. MacArthur, D.N. Johnstone, M. Sarahan, J. Taillon, T. Aarholt, pquinn-dls, V. Migunov, A. Eljarrat, J. Caron, C. Francis, T. Nemoto, T. Poon, S. Mazzucco, actions-user, N. Tappy, N. Cautaerts, S. Somnath, T. Slater, M. Walls, F. Winkler, H.W. Ånes, hyperspy/hyperspy: v2.0, (2023). https://doi.org/10.5281/zenodo.10412190.
[25] D. Johnstone, P. Crout, M. Nord, C. Francis, J. Laulainen, S. Høgås, E. Opheim, E. Prestat, B. Martineau, T. Bergh, N. Cautaerts, H.W. Ånes, S. Smeets, A. Ross, J. Broussard, S. Huang, S. Collins, T. Furnival, D. Jannis, I. Hjorth, E. Jacobsen, M. Danaie, A. Herzing, T. Poon, S. Dagenborg, R. Bjørge, A. Iqbal, J. Morzy, T. Doherty, T. Ostasevicius, T.I. Thorsen, M. von Lany, R. Tovey, P. Vacek, pyxem/pyxem: v0.17.0, (2024). https://doi.org/10.5281/zenodo.10551678.
[26] D. Johnstone, P. Crout, H.W. Ånes, E. Prestat, R. Tovey, S. Høgås, B. Martineau, J. Laulainen, N. Cautaerts, I. Wood, S. Collins, S. Smeets, A. Borrelli, T. Doherty, J. Morzy, E. Jacobsen, T. Bergh, T. Ostasevicius, E. Opheim, pyxem/diffsims: diffsims 0.5.2, (2023). https://doi.org/10.5281/zenodo.7962969.
[27] A. Valery, E.F. Rauch, L. ClÉment, F. Lorut, Retrieving overlapping crystals information from TEM nano-beam electron diffraction patterns, J Microsc 268 (2017) 208–218. https://doi.org/10.1111/jmi.12599.
[28] J.M. Gibson, M.M.J. Treacy, T. Sun, N.J. Zaluzec, Substantial crystalline topology in amorphous silicon, Phys Rev Lett 105 (2010) 1–4. https://doi.org/10.1103/PhysRevLett.105.125504.
[29] S. Huang, C. Francis, J. Sunderland, V. Jambur, I. Szlufarska, P.M. Voyles, Large Area, High Resolution Mapping of Approximate Rotational Symmetries in a Pd77.5Cu6Si16.5 Metallic Glass Thin Film, Ultramicroscopy 241 (2022). https://doi.org/10.1016/j.ultramic.2022.113612.