Microscopy and Microanalysis (2022), **28**, 371–382
doi:10.1017/S1431927622000010

Microscopy AND Microanalysis

Original Article

Cluster Analysis of Combined EDS and EBSD Data to Solve Ambiguous Phase Identifications

Chad M. Parish

Nuclear Energy Materials Microanalysis Group, Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Abstract

A common problem in analytical scanning electron microscopy (SEM) using electron backscatter diffraction (EBSD) is the differentiation of phases with distinct chemistry but the same or very similar crystal structure. X-ray energy dispersive spectroscopy (EDS) is useful to help differentiate these phases of similar crystal structures but different elemental makeups. However, open, automated, and unbiased methods of differentiating phases of similar EBSD responses based on their EDS response are lacking. This paper describes a simple data analytics-based method, using a combination of singular value decomposition and cluster analysis, to merge simultaneously acquired EDS + EBSD information and automatically determine phases from both their crystal and elemental data. I use hexagonal TiB_2 ceramic contaminated with multiple crystallographically ambiguous but chemically distinct cubic phases to illustrate the method. Code, in the form of a Python 3 Jupyter Notebook, and the necessary data to replicate the analysis are provided as Supplementary material.

Key words: EBSD, EDS, data analytics, data analysis

(Received 18 September 2021; revised 10 December 2021; accepted 3 January 2022)

Introduction

Overview

One of the most frustrating issues in electron backscatter diffraction (EBSD) is the inability of the analysis to easily differentiate between materials with the same or closely related crystal structures. For instance, nickel and copper are both $Fm\bar{3}m$ face-centered cubic crystals, with essentially the same lattice parameter ($a_0^{Ni} \approx 352$ pm, and $a_0^{Cu} \approx 361$ pm). Different methods already exist to differentiate phases in EBSD that cannot be differentiated by the simple Hough-based methods. For instance, dynamical pattern simulation followed by dictionary comparison (Chen et al., 2015) provides very high fidelity in indexing, but is computationally expensive and has significant start-up effort involved. Detector vendors have vendor-specific methods for merging the EBSD and energy dispersive spectroscopy (EDS) data to find the different phases. However, these methods can involve a certain amount of black box software and are also not transferrable from one vendor's system to another. These methods based on EDS elemental analysis (Nowell & Wright, 2004; Wright & Nowell, 2006; Nowell et al., 2011; Dietrich et al., 2014; Chiu et al., 2019) are effective, but they are *ad hoc*, dependent upon manually selected parameters, and are specific to the vendor of the detectors used in the experiment. Other methods (i.e., Bilsland et al. 2021) are very powerful, but solve more specific problems.

In this paper, I suggest an extension of the EDS-specific cluster analysis method of Stork & Keenan (2010) that merges the EDS-derived elemental information with the EBSD-derived crystallographic information. This provides an automated phase label for each pixel, with a minimum number of human-tuned parameters entering the analysis algorithm. It is also open-source and vendor-independent, only requiring the vendor to provide an open format export of the EDS and EBSD data [such as might be read by open-source packages like Hyperspy (de la Peña et al. 2017)].

Data Analytics in X-ray Spectrum Imaging

In an X-ray *spectrum image* (XSI), spatially resolved spectroscopic data is recorded such that a large number N of spectra are recorded at N discrete pixels, and each spectrum has M channels or data elements, leading to a large datacube of size (x pixels) × (y pixels) × M. To populate this datacube (Fig. 1), a beam is scanned across a sample [Fig. 1(a)], and the signal (such as X-rays) is collected by a detector, processed [Fig. 1(b)], and stored into computer memory as elements of a datacube [Fig. 1(c)]. Then, traditional X-ray maps can be extracted from the datacube as slices of energy space [Fig. 1(d)].

To analyze this datacube, the methods refined by Keenan et al. are used (Kotula et al., 2003; Keenan, 2007; Smentkowski et al., 2009). The XSI datacube is unfolded into a matrix **D** of size $N \times M$ and then analyzed [Fig. 1(e)]. *Singular value decomposition* (SVD) and its close cousin, *principal component analysis* (PCA), are methods of rank reduction and form the underlying mathematical framework to much of data science, and in the present

Cite this article: Parish CM (2022) Cluster Analysis of Combined EDS and EBSD Data to Solve Ambiguous Phase Identifications. *Microsc Microanal* **28**, 371–382. doi:10.1017/S1431927622000010

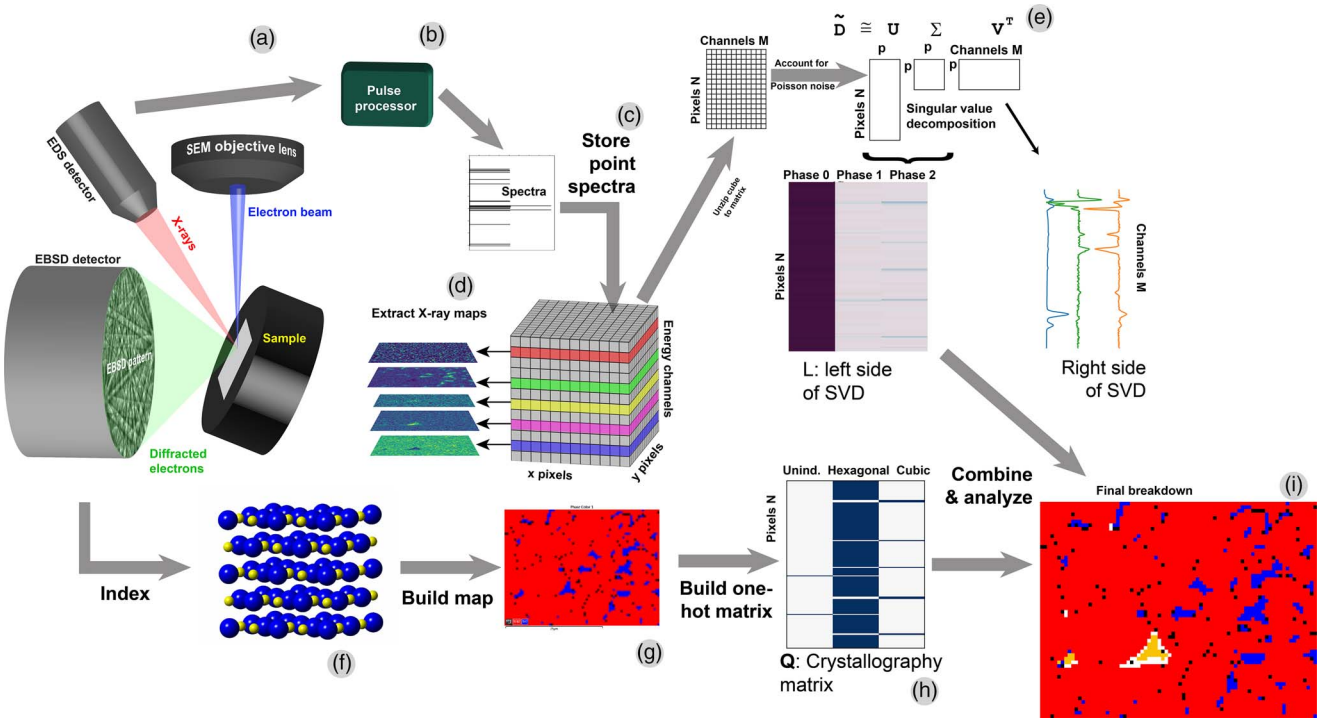

Fig. 1. Schematic of EDS + EBSD acquisition and analysis. (**a**) The EDS and EBSD data are acquired. (**b**) EDS pulses are converted and (**c**) stored into a datacube. (**d**) The traditional X-ray mapping functionality is extracted from this data. (**e**) The datacube is unfolded into a matrix and analyzed by singular value decomposition. (**f**) EBSD patterns are indexed to (**g**) build a phase map. This phase map is unfolded into a matrix and, when combined with the SVD information, can yield a combined chemical + crystallographic map (**i**).

context, form the backbone of data analytics methods used to analyze the matrix $\mathbf{D}$ of the XSIs.

Simultaneously to recording the XSI, the EBSD patterns are analyzed to find the crystal structure [Fig. 1(f)]. The array of EBSD pixels is combined into the phase map [Fig. 1(g)].

To perform SVD of X-ray SIs, it is important to use the insight of Keenan and Kotula (Kotula et al., 2003; Keenan et al., 2004*a*, 2004*b*), specifically that the count data $\mathbf{D}$ must be scaled to account for Poisson noise; this is because the noise between X-ray peaks and background are very different (heteroscedastic) and confound algorithms that assume uniform noise (homoscedastic). This is accomplished as described by Keenan and Kotula, specifically that the data matrix $\mathbf{D}$ is converted to scaled data $\tilde{\mathbf{D}}$ via:

$$\tilde{\mathbf{D}} = \mathbf{GDH}$$

This $\tilde{\mathbf{D}}$ is the "Poisson-scaled data matrix." As explained elsewhere (Keenan, 2007), $\mathbf{G}$ is a matrix of size $N \times N$ with the inverse square root of the mean image of $\mathbf{D}$ on the diagonal, and $\mathbf{H}$ is a matrix of size $M \times M$ with the inverse square root of the mean spectrum of $\mathbf{D}$ on the diagonal. This has the effect of making $\tilde{\mathbf{D}}$ very nearly homoscedastic, and therefore more amenable to analysis by SVD or PCA. Once scaled, the SVD is:

$$\tilde{\mathbf{D}} = \mathbf{U\Sigma V}^{\mathrm{T}}$$

The columns of $\mathbf{U}$ are the left singular vectors, the columns of $\mathbf{V}$ are the right singular vectors, superscript $\mathbf{T}$ denotes matrix transpose, and the diagonal of $\mathbf{\Sigma}$ carries the *singular values*. Importantly to the present discussion: (1) the columns of $\mathbf{V}$ are the chemical endmembers, and (2) the columns of $\mathbf{U}$ are, once refolded, the abundance maps associated with the chemical endmembers. The values of $\mathbf{\Sigma}$ are the relative strengths associated with each component. SVD requires that any two columns of $\mathbf{U}$ be mutually orthonormal, and similarly for columns of $\mathbf{V}$, and the singular values are sorted in descending order. As described elsewhere (Kotula et al., 2003; Burke et al., 2006; Watanabe et al., 2009; Parish, 2011), a "scree plot" of eigenvalues (singular values squared) can be plotted, and a knee or slope-break in the plot used to find the number of components to retain. Inverse transformation by $\mathbf{G^{-1}}$ and $\mathbf{H^{-1}}$ return the abundance maps and endmembers into real space from Poisson-scaled space (Keenan & Kotula, 2004*a*, 2004*b*), and post-processing tools such as varimax rotations (Keenan, 2005, 2009; Smentkowski et al., 2009) or independent component analysis (Hyvarinen, 1999; Windig & Keenan, 2015) are commonly applied to find more interpretable results.

Stork & Keenan (2010) and Keenan (2007) pointed out some disadvantages of the above-described matrix-decomposition methods. First, SVD and its ilk suffer a "parsimony restriction," in the terminology of Stork and Keenan. This means that the rank of the model cannot exceed the chemical rank of the sample. For instance, if only two chemical elements Fe and Cr are present, then the SVD model will have rank-2. However, if the material consists of, for example, elemental regions of Fe and Cr and an intermetallic compound FeCr, the rank-2-constrained SVD model cannot describe this three-region sample with a rank-3 model. This means that the FeCr region will be described by, for example, the Fe and Cr endmembers with roughly 50–50 abundances in the region of the compound. This requires the analyst to interpret this 50–50 mixture correctly, and the analyst will

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

not be presented with a clear FeCr endmember. The second issue with these methods is that "crisp" phase maps might not be achieved. PCA with varimax rotation to spatial simplicity (VRSS; Keenan 2009), ICA, etc., will provide realistic-looking endmembers, but give negative values in abundance maps (necessarily: if the average of a region's abundance map is to be zero, and there is noise, then some of the pixels must be <0). MCR-ALS (multivariate curve resolution-alternating least squares; Smentkowski et al., 2009) under non-negativity constraints, NNMF (non-negative matrix factorization), etc., can remove negative counts, but in turn result in errors in the endmember reconstructions, such as non-physical dips below the Bremsstrahlung in an endmember at the energy locations of peaks strongly present in other endmembers (e.g., Keenan 2009). Other methods, such as NFIND-R (Winter, 1999), are also in the early stages of exploration to apply to EDS.

It was the innovation of Stork & Keenan (2010), then, to suggest cluster analysis of EDS-SI data. *K*-means clustering is the classical example (Jain, 2010). In *K*-means, individual objects (pixels in an SI) are clustered based on their similarity; in this case, the similarity of their point spectra. An *a priori* number of clusters, *K*, is assigned by the analyst. *K*-means has the advantage that the cluster centers, which are the spectral endmembers, are very physical in appearance, without artifacts such as non-physical dips below the Bremsstrahlung. Furthermore, because the clusters need not be linearly independent, the parsimony restriction is lifted (Stork & Keenan, 2010). Highly efficient *K*-means packages are readily available (e.g., Python's `scikit-learn.cluster`; Pedregosa et al. 2011). However, *K*-means provides a hard assignment to individual pixels: a pixel is either in one cluster or another. Stork and Keenan suggested applying *fuzzy C-means clustering* (FCM), in which each pixel's assignment *sums* to 1.0, and a pixel can be shared among multiple clusters. This is vital, e.g., to describe mixed pixels at a phase boundary. Efficient and easy-to-use fuzzy *C*-means algorithms are available in Python as `scikit-fuzzy` (Warner et al., 2021).

Most importantly, Stork and Keenan found that clustering on the reduced left-side **L** of the Poisson-scaled SVD, where $\mathbf{L} = \mathbf{U\Sigma}$ from SVD of data $\tilde{\mathbf{D}}$ such that $\tilde{\mathbf{D}} = \mathbf{U\Sigma V}^{\mathrm{T}}$, is computationally very efficient. I illustrate this schematically in Figure 1e, which shows the unzipped matrix of spatial abundances (each row is a single pixel) and the spectral endmembers, for a rank $p = 3$ model. Furthermore, the final abundances' cluster centers (endmembers) are easily recovered from the reduced centers by premultiplication. The net result is to perform clustering on the small $N \times p$ matrix **L** instead of the large $N \times M$ matrix $\tilde{\mathbf{D}}$. Because $p << M$, this saves significant computational time. Clustering on the SVD reconstruction instead of the raw data also allows much improved analysis, because of SVD's noise filtering effects. Overall, the paper of Stork &Keenan (2010) should be read carefully for the details.

Application to EBSD

With those foundations laid, let's now discuss what is new in this paper. Specifically, let's define the hypothesis I will test:

Hypothesis: Combining EDS and EBSD information into a single matrix will allow efficient cluster analysis to separate crystallographically indistinguishable but chemically distinct phases.

I will take this one part at a time. First, scanning electron microscopy (SEM)-EDS by itself is known to be nicely amenable to cluster analysis, as discussed above in section "Data analytics in X-ray spectrum imaging," specifically by clustering on the reduced-rank SVD left-hand-side L.

The next—and more difficult—question is how to merge EDS and EBSD data into a format amenable to cluster analysis. Each pixel contains a spectrum of EDS data, but each pixel can contain one and only one crystallographic assignment. Therefore, I need a representation that can easily accommodate each pixel but multiple different crystallographic assignments. An initial thought might be to use an $N \times 1$ single-column matrix and assign each element in the matrix an integer value: 0 for unindexed, 1 for hexagonal, 2 for cubic, etc. However, cluster analysis can be sensitive to the absolute magnitude of different attributes (columns) of the matrix to be clustered. Which is to say, if a large number of crystals were present (10, perhaps), the clustering might give more "weight" to that column and less to the chemically derived columns of **L**.

To test my hypothesis, I therefore created a "one hot" matrix (Géron, 2019), so-called because each pixel has only a single nonzero ("hot") value. Specifically, I created a matrix, call it **Q**, where in its zeroth column, a "1" is assigned to each row (pixel) if the pixel is unindexed by EBSD; assigned a "1" in the first column of the pixel is indexed as hexagonal by EBSD; and assigned a "1" in the second column if the pixel is indexed as cubic (Fig. 1h). I use the convention that the first row, column, etc., is "zeroth" and counts up from zero, to be consistent with Python's array indexing. So, for N pixels and z number of crystallographic phases (in which the inevitable "unindexed" pixels are considered a distinct phase), **Q** is of size $N \times Z$. Because **L** is of size $N \times p$, then, a combined matrix **Y** of size $N \times (p + z)$ can be constructed as $\mathbf{Y} = [\boldsymbol{L}_{N\times p}\mathbf{Q}_{N\times z}]$; this concatenation is easily accomplished in Python via `numpy.hstack`. To account for the difference in absolute magnitudes of the values between **L** and **Q**, I divide **L** by the absolute value of the largest magnitude element in the matrix **L**, which reduces the maximum value of **L** to either +1 or −1, nicely matched to the one-hot values of exactly 1 in the matrix **Q**. To maintain the same model, **V** should then be multiplied by this same constant value. This yields, essentially, an arbitrary orthogonal factor model **LR** where **R** is this multiplied and then transposed **V**. My hypothesis, then, is that this combined matrix can be analyzed via clustering to provide a final chemistry and crystallography result that will differentiate chemically distinct but crystallographically indistinguishable phases (Fig. 1i).

In this paper, I will use example data of EDS + EBSD from a complex ceramic with crystallographically ambiguous phases to test this hypothesis.

Materials and Methods

Experimental

The material used was a TiB_2 ultra-high temperature ceramic, provided by the Missouri University of Science and Technology. The sample was prepared as described elsewhere (Bhattacharya et al., 2019). A small (several millimeters) piece was cut and polished to a mirror finish, with a final polishing step of 50 nm colloidal silica. This provided excellent EBSD patterns.

Simultaneous EDS and EBSD were acquired on a Tescan MIRA3 GMH field-emission SEM equipped with Oxford Instruments Symmetry CMOS-based EBSD detector and Oxford Instruments Ultim Max 170 mm^2 silicon-drift detector EDS. Data was acquired using Oxford Instruments AZtec 4.0 software. Data was acquired at 20 keV, 70° tilt, and ≈1 nA probe current. An 80 × 60 pixel, 40 × 30 μm map (500 nm pixel pitch) was

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

acquired. EBSD indexing was performed using TiB_2, space group 191 P6/mmm, and TiC, space group 225 $Fm\bar{3}m$, crystal cards. Importantly, TiO, TiC, TiN, and their alloy Ti(CNO) are entirely isostructural (with a rocksalt crystal structure) and indistinguishable crystallographically. The EDS data was exported in the Oxford .RAW (binary) / .RPL (text header) format, and the EBSD data exported in the Oxford .CTF (text) format.

In Figure 2, I show the EDS X-ray maps, EBSD phase and band contrast maps, and the forward-scattered electron image of the general region. X-ray maps are integrated net intensity windows whose full width is the estimated detector resolution at the energy E of the marked line, calculated by shifting 130 eV approximate resolution at Mn Kα (5,893 eV) via $\text{Width}(E) = [2.5(E\text{-}5{,}893) + (130^2)]^{1/2}$ (Goldstein et al., 1992).

Analysis

Anaconda Python 3 was used for all calculations. First, a custom script was used to read the EDS data from the .RAW/.RPL into memory, in which the X-ray point spectra and associated metadata were stored into numerical arrays. The CTF file containing the EBSD results listed each pixel's phase as "0" (unindexed), "1" (TiB_2 hexagonal), or "2" (TiC cubic). This was read into memory and converted into a NumPy array. This NumPy array was also converted into a "one-hot" matrix using Scikit-learn's OneHotEncoder. The OneHotEncoder produced a SciPy sparse matrix, and I then used its todense() method to convert the matrix from a sparse SciPy array to a dense NumPy array. This array's size was 4,800 × 3, for 4,800 pixels and 3 phases.

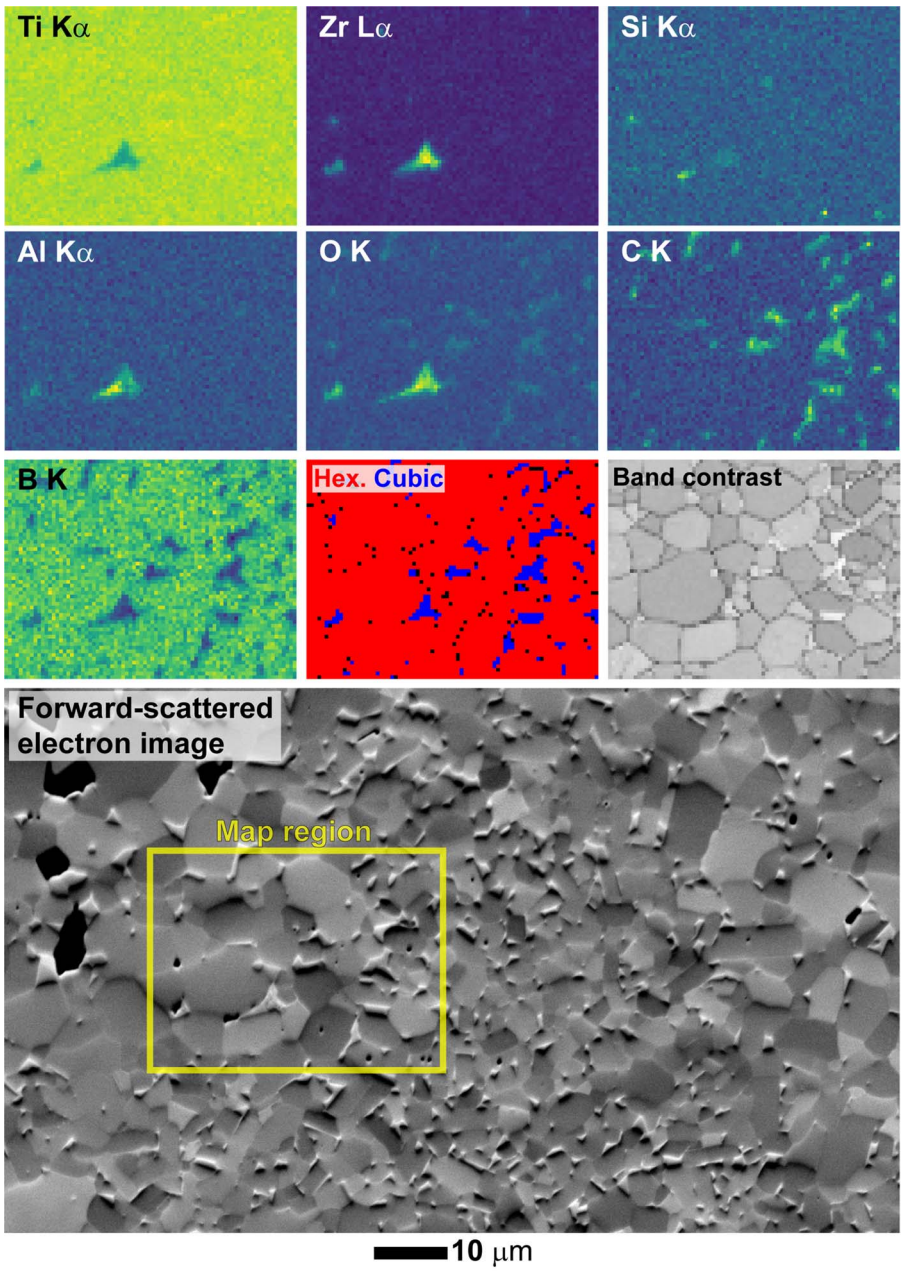

Fig. 2. X-ray maps for Ti, Zr, Si, Al, O, C, and B. The EBSD maps for phases (Hex. = hexagonal) and band contrast. The bottom image is the forward-scatter image, with the map region marked. All figures have the same scale.

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

The EDS data was trimmed to −0.2 to 5.79 keV (the Oxford electronics giving some empty channels at negative energies), converted from a dense NumPy array into a compressed sparse column SciPy array, and then scaled for Poisson noise as described above. Because of the high beam current and long pixel dwell time needed for EBSD, the EDS signal was strong and binning was not needed to extract meaningful singular values. SVD was performed on the Poisson-scaled sparse matrix using Python: scipy.sparse.linalg.svds (Virtanen et al., 2020). The SVD yields the triplet $\hat{\mathbf{D}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$, and I then define $\mathbf{L} = \mathbf{U}\boldsymbol{\Sigma}$. Recall that $\mathbf{L}$ is size $N \times p$, where N is the number of pixels ($80 \times 60 = 4{,}800$) and p is the rank of the SVD, which is this case is $p = 3$.

Results and Discussion

EBSD finds two crystallographic phases, but EDS finds three chemical phases. *This is, of course, the crux of the problem I wish to address.* The phase map in Figure 2 labels these two crystallographic phases as "Hex." (Hexagonal, red) and "Cubic" (blue). Unindexed pixels, essentially a third null phase, are black.

The PCA analysis, Figure 3, is the result of returning the noise-scaled SVD analysis from the Poisson-scaled space back to the original real space, and then re-orthogonalizing using Keenan's "fPCA" method (Keenan, 2007). The first component (Fig. 3, top row) is the mean spectrum and mean image. The second component (Fig. 3, middle row) shows the differences between the matrix and the precipitates. The Ti and B peaks are negative in that endmember, which indicates the precipitates are weak in these elements. These negative "counts" are indicative of the problem with simple SVD/PCA analyses: their difficult interpretability. The third component (Fig. 3, bottom row) shows the differences between the two precipitate populations.

A more interpretable approach is seen in Figure 4, which is the "varimax rotation to spatial simplicity" (VRSS) (Keenan, 2009) analysis. This shows that there are three distinct regions, specifically, TiB_2 matrix (top row); (Ti,Zr,Al,O,(Ca?))-oxide precipitates (middle row); and Ti(C,N,O) precipitates (bottom row). Other methods, such as varimax-rotated PCA, NNMF, and NFIND-R, are certainly possible and give substantially similar results, but VRSS provides a clear interpretation of the X-ray data here.

From this starting point, I begin the specific analysis of the data to explore the use of combined elemental and crystallographic information via clustering.

First is *K*-means and fuzzy *C*-means analysis of the EDS data. *K*-means cluster analysis of the left-side of the SVD, **L**, is shown in Figure 5 and fuzzy *C*-means in Figure 6. These use the methodology

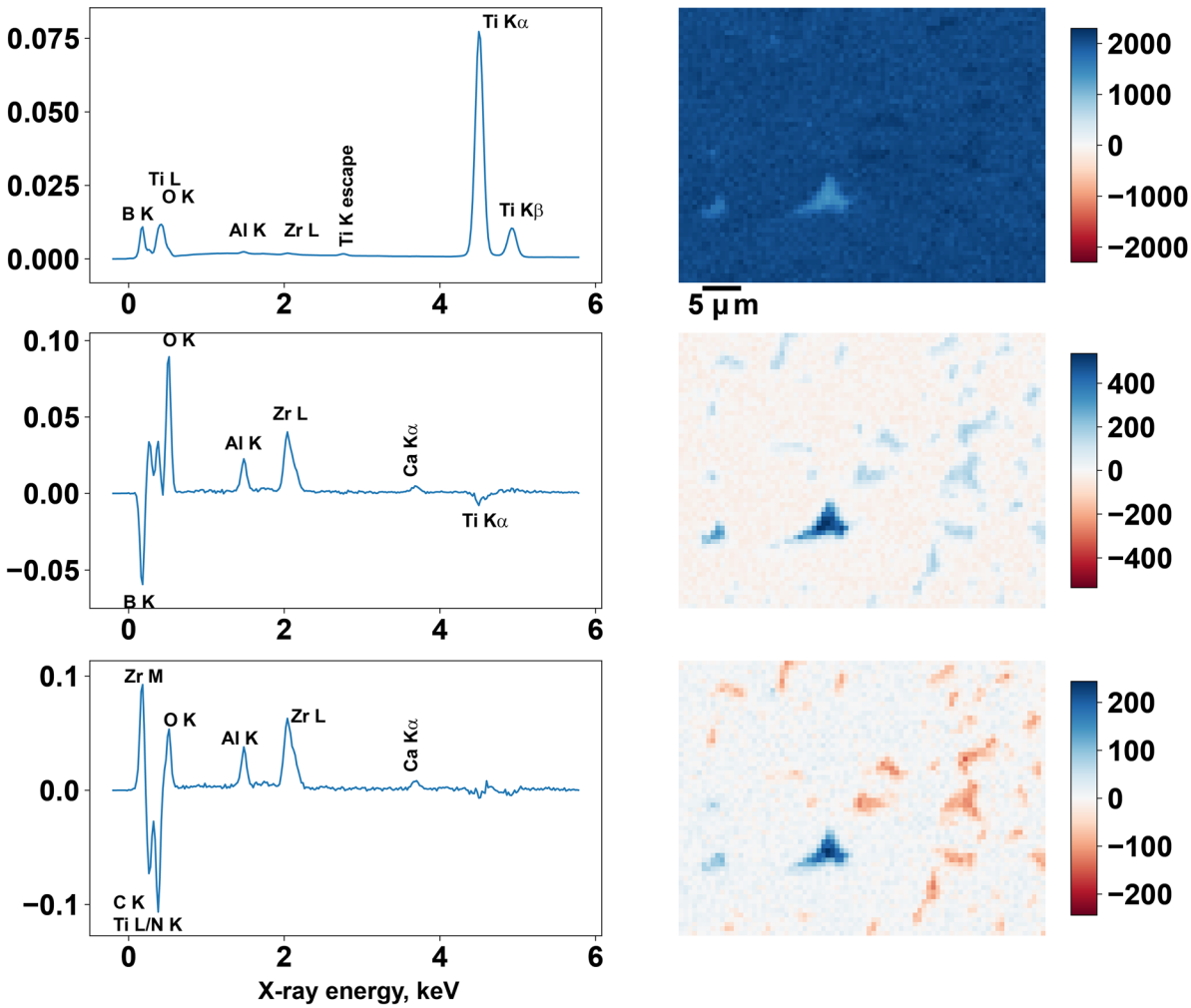

Fig. 3. PCA of the XSI, returned from noise-scaled space to real space and re-orthogonalized. The first component (top row) describes the mean (primarily Ti-Zr-Al-O-B). The second component describes the differences between the matrix and the precipitates. The third component (bottom row) describes the differences between the precipitates. Spectral endmembers are scaled to an integral of 1.0.

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

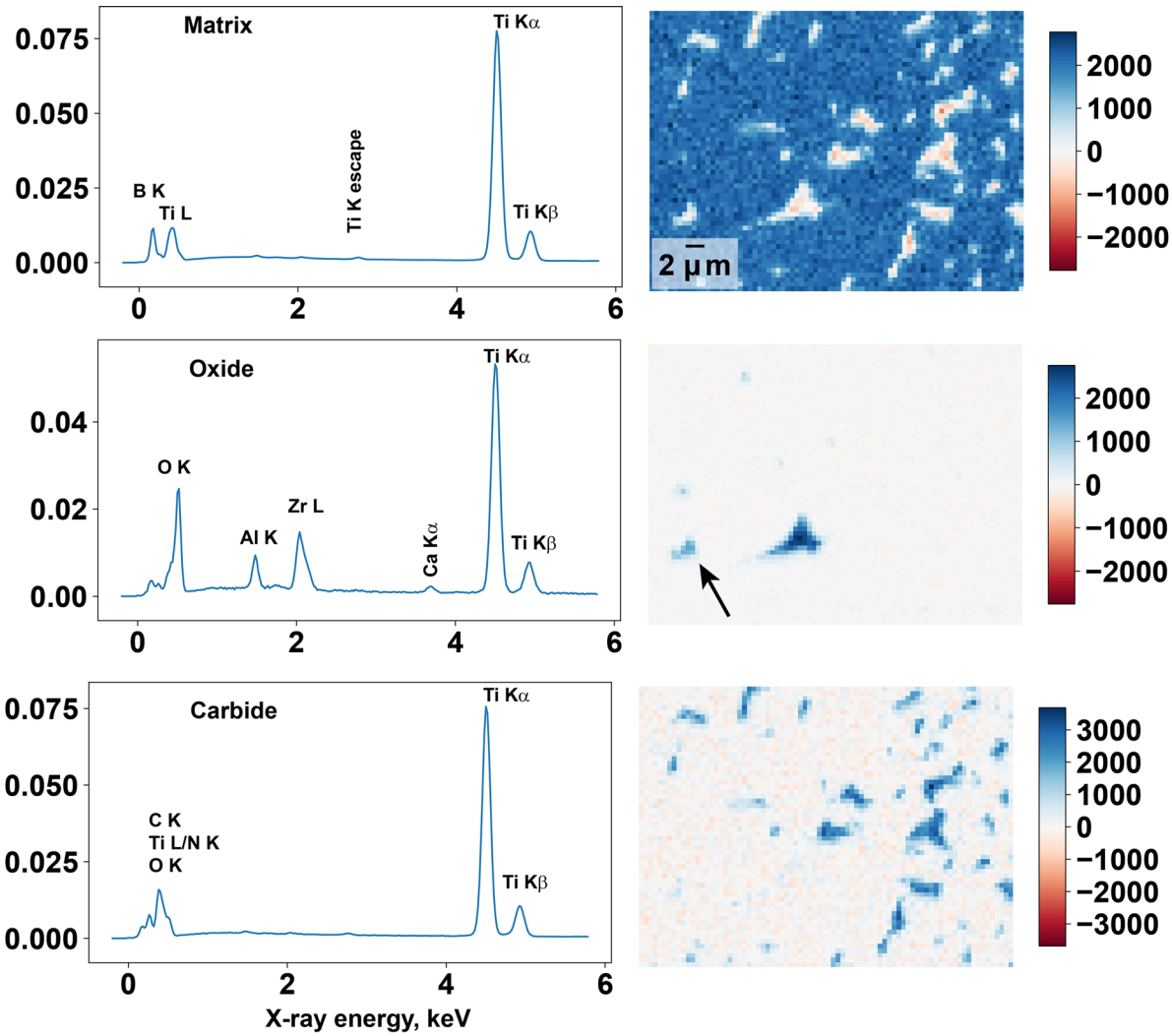

Fig. 4. Varimax rotation to spatial simplicity analysis of the XSI. Three independent components are found: Ti-B (top); (Ti,Zr,Al,O,(Ca?)) (middle); Ti-(C,N,O) (bottom). Spectral endmembers are scaled to an integral of 1.0. The arrow indicates a precipitate of interest in later analysis.

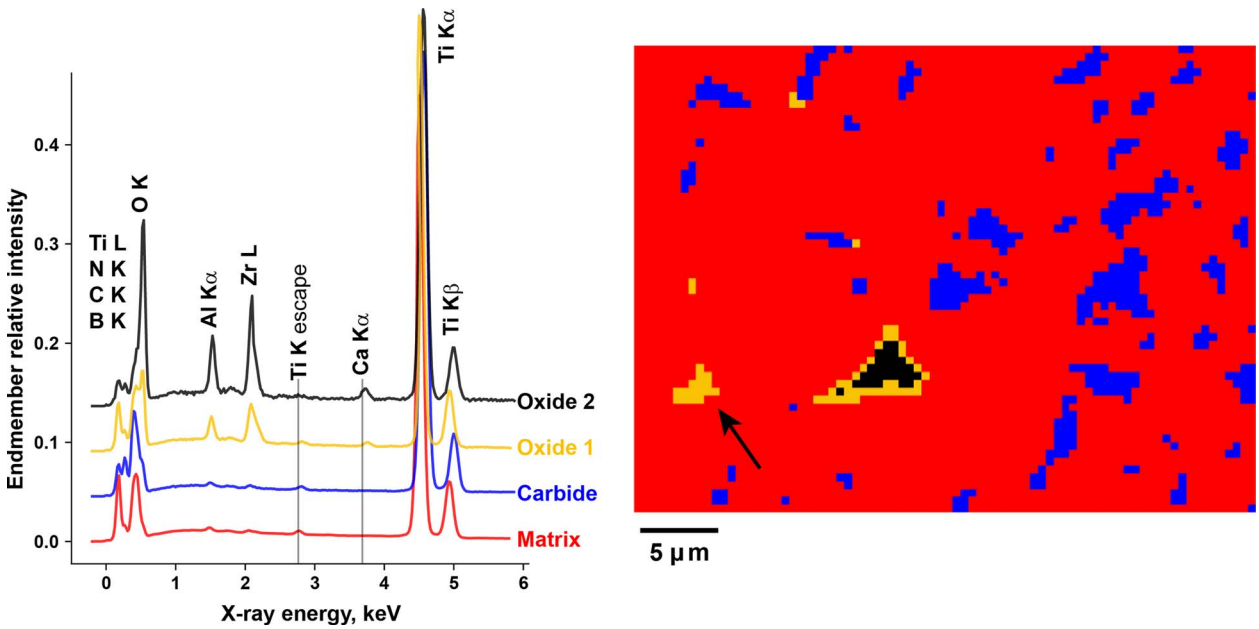

Fig. 5. *K*-means clustering analysis on the SVD-reduced EDS data. Four clusters (phases) describe the data very well. Spectral endmembers are arbitrarily scaled and offset vertically for visibility.

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

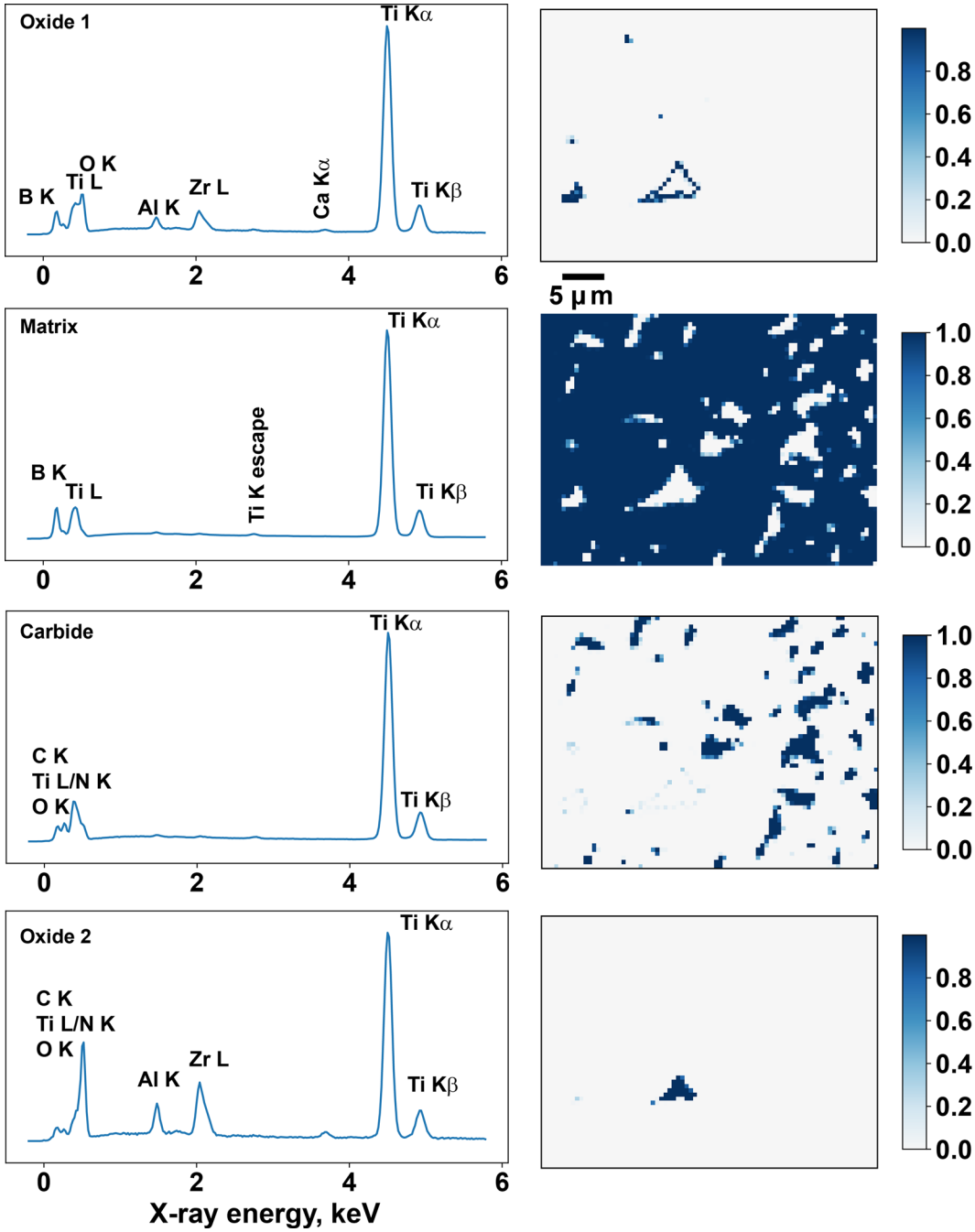

Fig. 6. Fuzzy *C*-means clustering of the SVD-reduced EDS data. Four clusters (phases) describe the data very well. Spectral endmembers are arbitrarily scaled.

described for fuzzy *C*-means described by Stork & Keenan (2010). *K*-means was calculated using scikit-learn (Pedregosa et al., 2011) and fuzzy *C*-means using scikit-fuzzy (Warner et al. 2021). From this starting point, I begin the specific analysis of the data to explore the use of combined elemental and crystallographic information via clustering. For both cluster analyses, I found that four clusters provided the most insightful analysis. With three clusters, the result replicates the VRSS results. With four clusters, an additional component not found in the VRSS result is discovered, providing new insight. However, five clusters resulted in non-physical checkerboarding in which the matrix (TiB_2 phase) was split into two different clusters without any physical meaning.

The reason that four clusters, instead of three, appears more insightful can be explained by looking at the precipitate that is arrowed in Figures 4–6. In either cluster analysis figure, the "oxide 1" spectral endmember shows more B-K X-ray intensity and less O-K X-ray intensity than "oxide 2"; oxide 1 looks like a linear combination of the matrix and oxide 2, and this point will come back later. In the VRSS analysis (Fig. 4), the same particle shows weaker "oxide" contribution than the large oxide particle, and more contribution from the matrix phase. This is likely interpretable as the arrowed particle being very thin, and the 20 kV electron beam exciting noticeable amounts of B, and noticeably less O, at those points. This illustrates nicely Stork

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

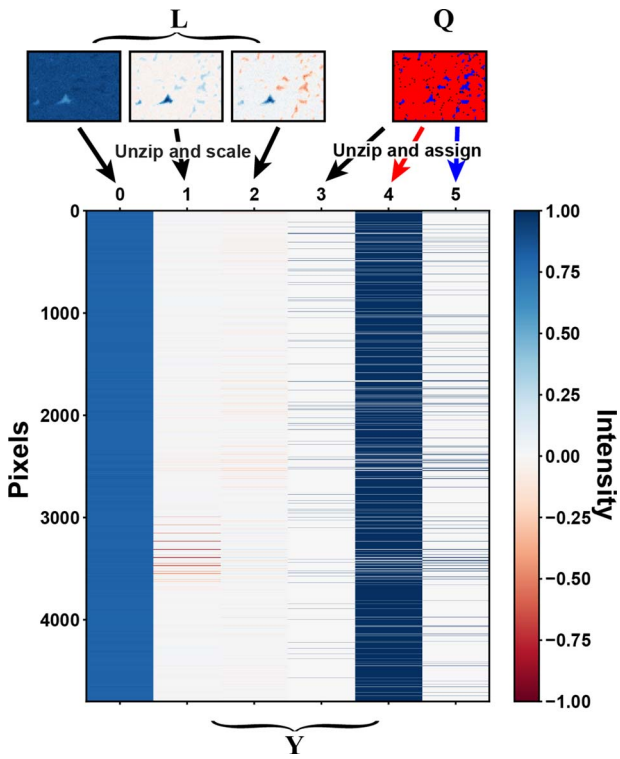

Fig. 7. Illustration of unzipping L and Q into a combined matrix Y for analysis. Note that L is scaled to a maximum absolute value of 1.0.

and Keenan's point that cluster analysis does not suffer the "parsimony restriction" that factor analysis methods suffer (Stork & Keenan, 2010).

So, now, the next question is, how can the chemical (EDS) and crystallographic (EBSD) information be combined into a single comprehensive analytical look at the sample? As noted above in section "Application to EBSD," I hypothesize that a cluster analysis applied to a combined matrix of chemical information and crystallographic information will provide the desired differentiation of crystallographically indistinguishable but chemically distinct phases.

At first, I can combine the SVD left-side **L** (scaled to a maximum absolute value of 1.0) and a matrix **Q** of the crystallographic identifications into a single matrix **Y**. This is illustrated in Figure 7. Then, *K*-means clustering of this matrix **Y** is performed. *K*-means clustering, as in Figure 5, produces a phase assignment map and spectral endmembers. The cluster centers of a $K = 5$ analysis of the matrix in Figure 7 is given in Figure 8a. Figure 8a shows the cluster centers directly derived from clustering Figure 7. Figure 8b shows the cluster assignment map. Although Stork & Keenan (2010) give an equation that allows the recovery of the (Poisson-space) endmembers, **P**, from the SVD's right-side endmembers, $\mathbf{V_k}$ and the cluster centers $\hat{\mathbf{P}}$, such that $\mathbf{P} = \mathbf{V}_k\hat{\mathbf{P}}$, and which can be returned to real-space by dividing by the spectral Poisson-scaling vector **H**, here I found the simply summing the point spectra under each cluster map provides an identical result (±numerical rounding after normalization to highest peaks), and use the summed spectra under each cluster assignment map, scaled to the number of pixels per phase as that endmember (i.e., cluster centers). It is important to note that the final three columns of the cluster centers are the crystallographic assignments and should be truncated before applying that equation. The final cluster centers, in real (spectral) space, are given in Figure 8c and a detailed view of the low energy lines in Figure 8d.

Figures 8c and 8d show the insight gained from the proposed method. Cluster #0 (red) is the matrix and shows Ti and B lines clearly and was assigned to the hexagonal phase. So far, so good: this is the TiB_2 matrix. Cluster #1 (blue) shows Ti and C and a cubic assignment. This, also, makes sense given the X-ray maps and EBSD maps in Figure 2; this is a carbon-rich Ti(C,N,O) alloy and will be called TiC for simplicity. Cluster #2 (black) is chemically indistinguishable from the matrix but is "unindexed" phase assignment; these are matrix pixels with bad EBSD patterns, mostly from grain boundaries that gave overlapped Kikuchi bands. This phase could have been eliminated by cleaning the EBSD results (Brewer & Michael, 2010) but I decided not to apply EBSD data cleaning on my data.

It is for Cluster #3 (gold) where the benefit of this new technique becomes apparent. The EBSD map in Figure 2 showed the triangular particle on the left-center of the map as cubic, and indistinguishable from the carbides of Cluster #2. However, here, the chemistry seen in the spectral endmember is entirely different—Al,Zr,Ti,O(,Ca)—compared to the TiC chemistry of Cluster #2 and despite the cubic phase assignment that, by EBSD alone, was indistinguishable from the TiC. *Thus, the hypothesis in* section *"Application to EBSD" is confirmed: combined EDS + EBSD via cluster analysis has indeed differentiated chemically distinct but crystallographically ambiguous phases*. Had I indexed with $K = 4$, this would be a satisfying end to the analysis. However, the $K = 5$ clustering shows the fifth cluster, Cluster #4 (gray), which is very surprising indeed. The Al and Zr lines are intermediate between the Al,Zr,Ti,O(,Ca) (gold) and TiB_2 (red) phases, as is the B line. The phase assignment is, surprisingly, hexagonal. This phase, then, appears to identify TiB_2 (hexagonal) pixels in which the blooming of the electron beam underneath the surface excited X-rays from a region larger than the small surface area sampled by EBSD. Cluster #4 (gray), then, describes the different spatial resolutions between the EDS and EBSD signals. Indeed, a Monte Carlo simulation using CASINO (Hovington et al., 1997) shows ~1,000 nm beam broadening of a 20 keV beam at 70° tilt onto TiB_2 of 4.5 g/cm^3; compares this roughly 1,000 nm to the 500 nm step size of the experiment and the very small (<10 nm) size of the actual field emission probe at a modest beam current of ≈1 nA. Although EBSD spatial resolution is difficult to estimate, it will be closer to the probe size than the interaction volume; based on measurements on other materials (Chen et al., 2011), around 100 nm seems a reasonable guess here.

It is a well-known limitation of clustering methods that if different features being clustered on very different scales, the results of the cluster analysis can be skewed; therefore, I will try the same analysis, but instead of using **L** (where $\mathbf{L} = \mathbf{U\Sigma}$), what if only **U** is used as the basis for the clustering? In this analysis, the trace of **Σ** is [1,200.0, 118.2, 71.5]. Which is to say, the 0th column of **L** is about 10× stronger than the 1st column, which is about twice as strong as the 2nd column. **U**, conversely, is orthonormal (each column is a unit vector), so the scales are much closer. Figure 9a shows this matrix; a comparison of Figure 9a to Figure 7 shows that the absolute values of the elements in columns 0, 1, and 2 are much closer. It is therefore reasonable to assume the clustering behavior will be more sensitive to smaller features. The actual cluster centers in Figure 9b are not appreciably different from those in Figure 8a—note that the order in

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

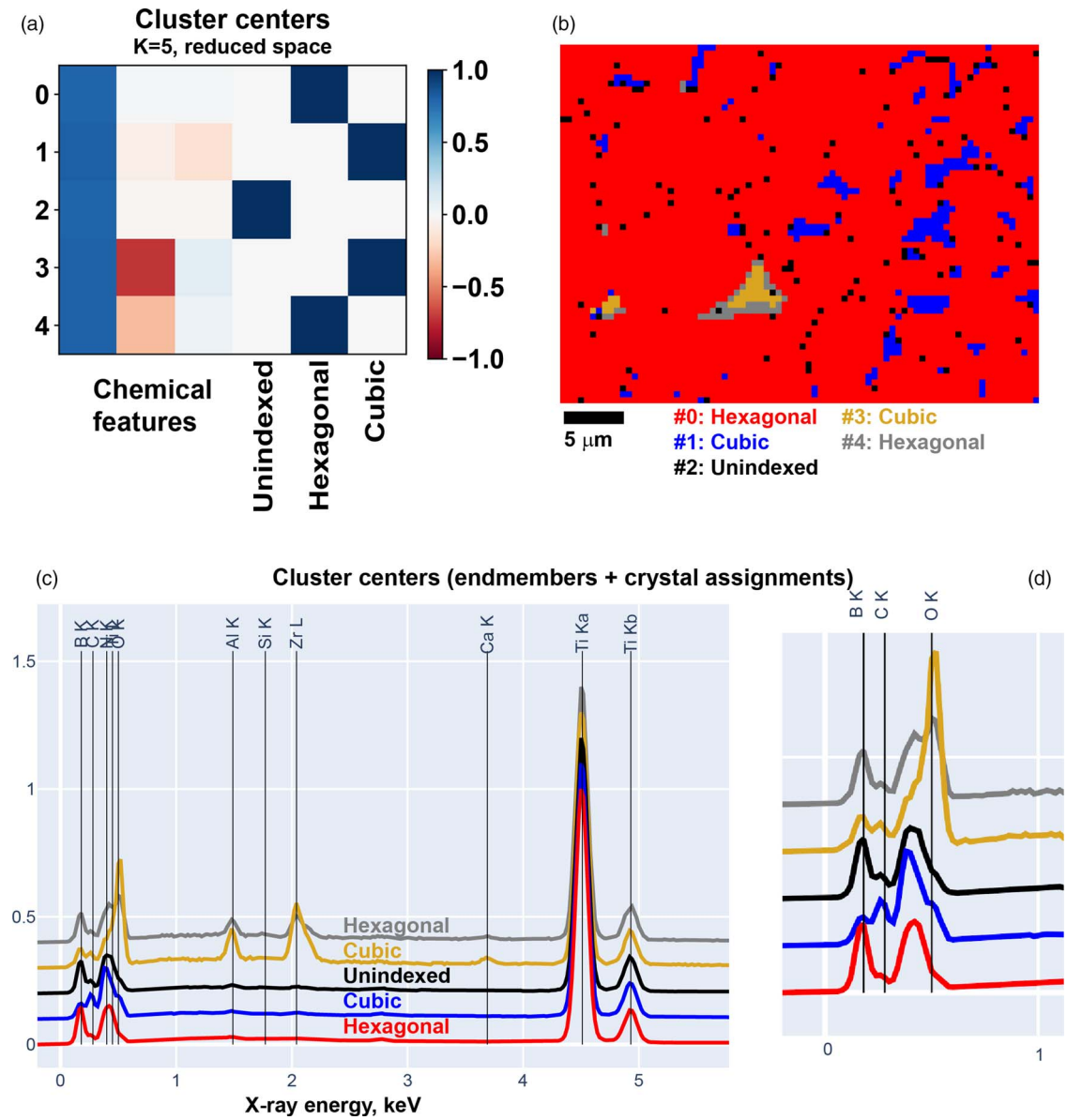

Fig. 8. (**a**) Cluster centers found by $K = 5$ *K*-means of the matrix in Figure 7. (**b**) The associated cluster assignment map. (**c**) The cluster centers after conversion from the reduced SVD space to real space. (**d**) A detailed view of the low-energy X-ray lines.

which clusters are assigned is arbitrary. It is, in Figure 9c, the cluster assignment map, where this **U**-based analysis begins to differ from the **L**-based analysis. The assignment is a bit noisier; speckle is observed, which was absent from the **L**-based cluster assignment map, Figure 8b. The Al,Zr,Ti,O(,Ca) cubic phase (gold), the TiB_2 hexagonal matrix phase (red), and the unindexed TiB_2 phase (black) are unchanged. Careful inspection of the carbide cubic phase (blue) finds little or no difference between the **L**-based and **U**-based assignment, but large differences appear in the chemically mixed hexagonal phase (gray). First, gray edging is seen around the carbides in the **U**-based which was not observed in the **L**-based. These seem to indicate that the threshold to assign the phase's chemistry to the gray phase is lower in the **U**-based assignment. This makes sense because the scale difference between the phases was removed and smaller effects in the non-matrix should be visible. More interesting, additional entire regions, such as the gray particle marked by the white arrow in Figure 9c, are discovered. Careful inspection of the X-ray maps (Fig. 2) indicates that, indeed, a *very weak* O and C signal is seen at that location. Given the hexagonal indexing and lack of visible boundary at that site in the band contrast map, this may be a shallowly buried oxycarbide excited by the penetrative 20 kV beam.

Ultimately, the comparison between the **U**-based (Fig. 9) and **L**-based (Fig. 8) analyses appears to be that the **U**-based is more sensitive to both noise and small signals, so the choice of one over the other might involve a decision toward sensitivity versus map crispness; for these maps, the *K*-means computation times were <5 s, so analyzing in both methods and comparing them are resource-light and thorough.

For one last comparison, the fuzzy *C*-means results (Fig. 6) were converted into an **L**-like matrix, where if a pixel was >0.8 in a single cluster, it was assigned 100% to that pixel, and the same clustering performed. Initial results were grossly incorrect,

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

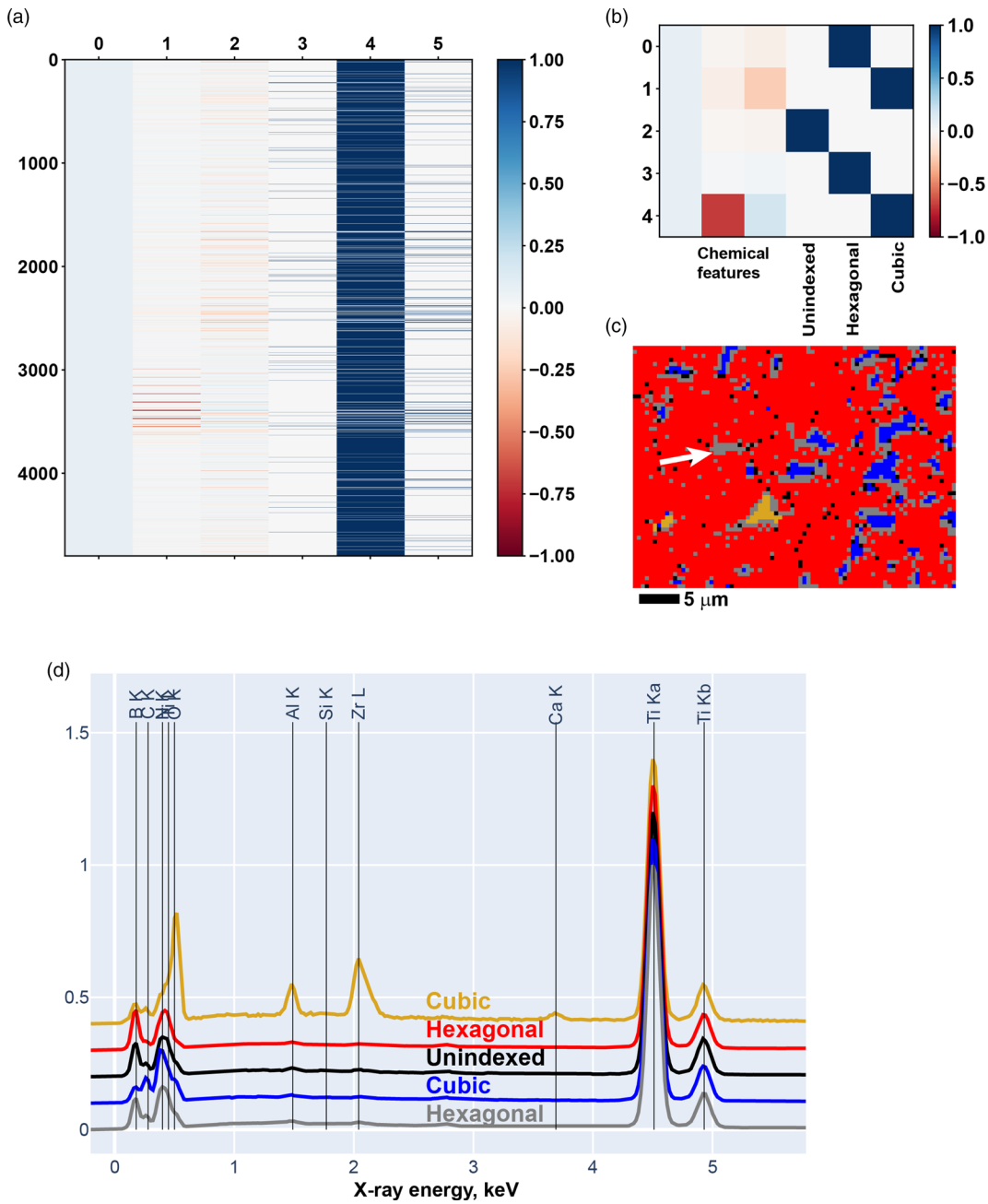

Fig. 9. (**a**) U-based construction of the chemistry + crystallography matrix Q. (**b**) Cluster centers obtained by $K = 5$ K-means of the matrix in (**a**). (**c**) Cluster assignment map. (**d**) Chemistry endmembers + phase assignments derived from (**b**).

in that with $K = 5$, the TiC and Al,Zr,Ti,O(,Ca) cubic phases were assigned to the same cluster. However, if analyzed with $K = 6$, this FCM-based analysis provides a very similar result to the **U**-based analysis, Figure 10. The TiB_2 (red), Al,Ti,Zr,O(,Ca) (gold), and TiC (blue) phases are unchanged. The hexagonal edging-effect phase (gray) is seen again, including at the "buried" feature seen in Figure 9c arrowed, but a sixth assignment, very similar to the gray assignment, green in Figure 10, shows a second edge-effect hexagonal phase with perhaps very slightly less of an O-line than the gray edge-effect phase. In other words, this provides perhaps the cleanest look yet—the buried phase is found like in the **U**-based but with less speckle—although at the cost of slightly more complicated interpretation, owing to the sixth phase.

Conclusions

The proposed cluster analysis method reliably and effectively differentiated the two cubic phases, TiC and Al,Zr,Ti,O(,Ca), in this tested sample. Different methods already exist to differentiate phases in EBSD that cannot be differentiated by the simple Hough-based methods. For instance, dynamical pattern simulation followed by dictionary comparison (Chen et al., 2015) provides very high fidelity in indexing, but is computationally expensive and has significant start-up effort involved. Detector vendors have vendor-specific methods for merging the EBSD and EDS data to find the different phases. However, these methods can involve a certain amount of

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

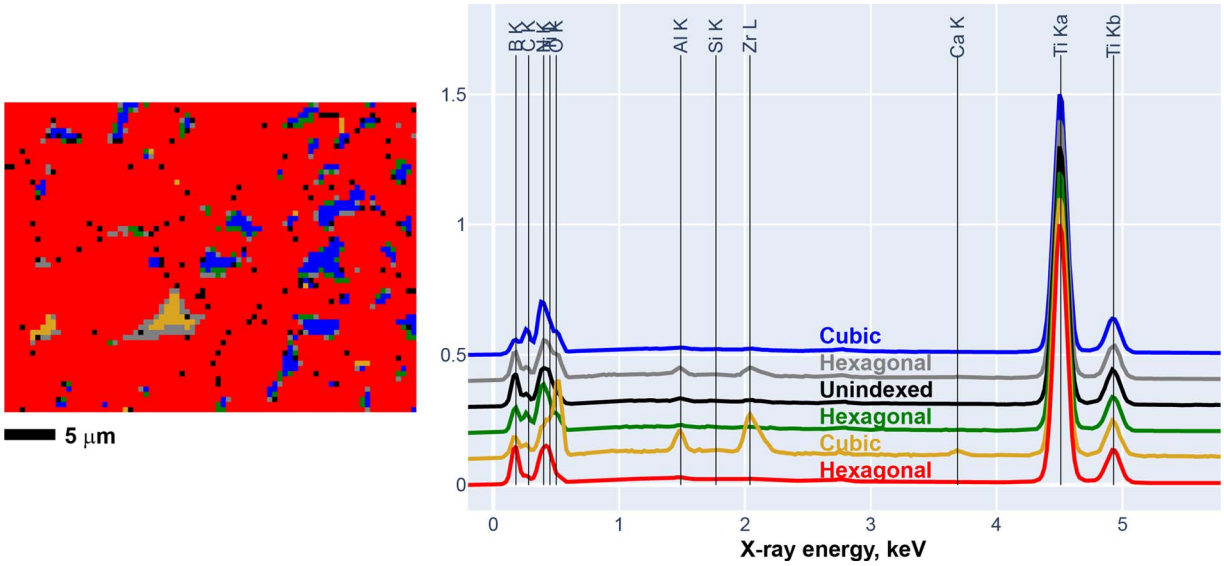

Fig. 10. Cluster assignment maps and cluster-center endmembers + crystallographic assignments for the FCM-based analysis. The "cubic" phases are TiC (blue) and (Al, Zr, Ti, O) (gold). The hexagonal phases are TiB_2 (red), Ti-B-O (Green), and Al-Zr-Ti-B mixed (gray).

black-box software and are also not transferrable from one vendor's system to another.

The proposed method has several advantages:

1. The proposed method successfully differentiates chemically different phases that had indexed identically using the on-microscope Hough-based approach.
2. The proposed method is computationally cheap; for the 80 × 60 pixel dataset here, the full analysis is in the order of minutes, and the majority of that time is spent on the FCM clustering (Fig. 6), because the FCM is repeated 512 times to help ensure convergence to an optimum solution from the random initializers.
3. This method is vendor agnostic, and only requires that the EBSD phase assignments and the EDS pixel spectra be readable and tagged to the pixels.
4. This method is provided in the Supplementary material, section 6, as open-source software and is freely available.

The primary disadvantages to the present technique are the following:

1. The choice of the number of clusters to determine requires a combination of domain knowledge and trial and error; however, the vendor-specific methods share this drawback.
2. Vendor-specific methods can be performed using the on-microscope software and do not require any programming knowledge.
3. The choice of **L**-based, **U**-based, or FCM-based starting matrix provides somewhat different results; however, the major phase assignments are unchanged, and only the very small differences (such as tiny contributions of Al-Zr to the matrix spectral component) are different.

Regardless, it seems that this method has potential as a machine-learning approach to the differentiation of crystallographically ambiguous phases via combined EDS + EBSD. Further development is promising as a simple and effective way to identify such phases.

Acknowledgments. The author thanks to Dr. Hanns Gietl and Dr. John Echols, ORNL, for critiquing the manuscript. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). Sample courtesy G. Hilmas & W. Fahrenholtz, Missouri University of Science and Technology.

Financial support. This work was supported by US Department of Energy, Office of Science, Fusion Energy Sciences, under contract number DE-AC05-00OR22725.

References

Bhattacharya A, Parish CM, Koyanagi T, Petrie CM, King D, Hilmas G, Fahrenholtz WG, Zinkle SJ & Katoh Y (2019). Nano-scale microstructure damage by neutron irradiations in a novel Boron-11 enriched TiB_2 ultra-high temperature ceramic. *Acta Mater* **165**, 26–39.

Bilsland C, Barrow A & Britton TB (2021). Correlative statistical microstructural assessment of precipitates and their distribution, with simultaneous electron backscatter diffraction and energy dispersive X-ray spectroscopy. *Mater Charact* **176**, 111071.

Brewer LN & Michael JR (2010). Risks of "cleaning" electron backscatter diffraction data. *Micros Today* **18**(2), 10–15.

Burke MG, Watanabe M, Williams DB & Hyde JM (2006). Quantitative characterization of nanoprecipitates in irradiated low-alloy steels: Advances in the application of FEG-STEM quantitative microanalysis to real materials. *J Mater Sci* **41**(14), 4512–4522.

Chen D, Kuo J-C & Wu W-T (2011). Effect of microscopic parameters on EBSD spatial resolution. *Ultramicroscopy* **111**(9–10), 1488–1494.

Chen YH, Park SU, Wei D, Newstadt G, Jackson MA, Simmons JP, De Graef M & Hero AO (2015). A dictionary approach to electron backscatter diffraction indexing. *Microsc Microanal* **21**(3), 739–752. doi:10.1017/S1431927615000756

Chiu Y, Yu H, Hung H, Wang Y & Kao C (2019). Phase formation and microstructure evolution in Cu/In/Cu joints. *Microelectron Reliab* **95**, 18–27.

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press

de la Peña F, Ostasevicius T, Fauske VT, Burdet P, Jokubauskas P, Nord M, Sarahan M, Prestat E, Johnstone DN & Taillon J (2017). Electron microscopy (big and small) data analysis with the open source software package HyperSpy. *Microsc Microanal* **23**(S1), 214–215. doi:10.1017/S1431927617001751

Dietrich D, Grittner N, Mehner T, Nickel D, Schaper M, Maier H & Lampke T (2014). Microstructural evolution in the bonding zones of co-extruded aluminium/titanium. *J Mater Sci* **49**(6), 2442–2455.

Géron A (2019). *Hands-On Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media.

Goldstein JI, Newbury DE, Echlin P, Joy DC, Romig AD Jr, Lyman CE, Fiori C & Lifshin E (1992). *Scanning Electron Microscopy and X-ray Microanalysis*, 2nd ed. New York: Plenum.

Hovington P, Drouin D & Gauvin R (1997). CASINO: A new Monte Carlo code in C language for electron beam interaction—Part I: Description of the program. *Scanning* **19**(1), 1–14.

Hyvarinen A (1999). Fast ICA for noisy data using Gaussian moments. In *1999 IEEE international symposium on circuits and systems (ISCAS)*. IEEE.

Jain AK (2010). Data clustering: 50 years beyond K-means. *Pattern Recognit Lett* **31**(8), 651–666.

Keenan MR (2005). Maximum likelihood principal component analysis of time-of-flight secondary ion mass spectrometry spectral images. *J Vac Sci Technol A* **23**(4), 746–750.

Keenan MR (2007). Multivariate analysis of spectral images composed of count data. In *Techniques and Applications of Hyperspectral Image Analysis*, Grahn HF & Geladi P (Eds.), pp. 89–126. Chichester: John Wiley & Sons.

Keenan MR (2009). Exploiting spatial-domain simplicity in spectral image analysis. *Surf Interface Anal* **41**, 79–87.

Keenan MR & Kotula PG (2004*a*). Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf Interface Anal* **36**(3), 203–212.

Keenan MR & Kotula PG (2004*b*). Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis. *Appl Surf Sci* **231-232**, 240–244.

Kotula PG, Keenan MR & Michael JR (2003). Automated analysis of SEM X-ray spectral images: A powerful new microanalysis tool. *Microsc Microanal* **9**(1), 1–17.

Nowell MM, Anderhalt R, Nylese T, Eggert F, de Kloe R, Schleifer M & Wright SI (2011). Improved EDS performance and EBSD geometry. *Microsc Microanal* **17**(Suppl. 2), 298–299.

Nowell MM & Wright SI (2004). Phase differentiation via combined EBSD and XEDS. *J Microsc* **213**, 296–305.

Parish CM (2011). Multivariate statistics applications in scanning transmission electron microscopy X-ray spectrum imaging. *Adv Imaging Electron Phys* **168**, 249–295.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R & Dubourg V (2011). Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825–2830.

Smentkowski VS, Ostrowski SG & Keenan MR (2009). A comparison of multivariate statistical analysis protocols for ToF-SIMS spectral images. *Surf Interface Anal* **41**, 88–96.

Stork CL & Keenan MR (2010). Advantages of clustering in the phase classification of hyperspectral materials images. *Microsc Microanal* **16**, 810–820.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W & Bright J (2020). Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* **17**(3), 261–272.

Warner J, Sexauer J, Unnikrishnan A, Castelão G, Pontes FA, Uelwer T, Batista F, Van den Broeck W, Song W, Martínez Pérez RA, Power JF, Mishra H, Trullols GO & Hörteborn A (2021). "Scikit-fuzzy." Available at https://zenodo.org/record/3541386.

Watanabe M, Okunishi E & Ishizuka K (2009). Analysis of spectrum-imaging datasets in atomic-resolution electron microscopy. *Microsc Anal* **23**(7), 5–7.

Windig W & Keenan M (2015). Homeopathic ICA: A simple approach to expand the use of independent component analysis (ICA). *Chemom Intell Lab Syst* **142**, 54–63.

Winter ME (1999). N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. *Imaging Spectrometry V. International Society for Optics and Photonics* **3753**, 266–275.

Wright SI & Nowell MM (2006). EDS assisted phase differentiation in orientation imaging microscopy. In *Advanced Structural Materials II*, vol. 509, Balmori-Ramirez H, Brito M, Cabanas-Moreno JG, Calderón-Benavides HA, Ishizaki K & Salinas-Rodríguez A (Eds.), pp. 11–16.

https://doi.org/10.1017/S1431927622000010 Published online by Cambridge University Press