UNIVERSITY of York

This is a repository copy of *Correlated electron diffraction and energy-dispersive X-ray for automated microstructure analysis*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/201790/

Version: Published Version

Article:
Duran, E. C., Kho, Z., Einsle, J. F. et al. (5 more authors) (2023) Correlated electron diffraction and energy-dispersive X-ray for automated microstructure analysis. Computational Materials Science. 112336. ISSN 0927-0256

https://doi.org/10.1016/j.commatsci.2023.112336

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

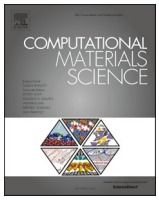
eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Computational Materials Science 228 (2023) 112336

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci



Full Length Article

Correlated electron diffraction and energy-dispersive X-ray for automated microstructure analysis

E.C. Duran [a], Z. Kho [a], J.F. Einsle [b], I. Azaceta [c], S.A. Cavill [c], A. Kerrigan [d], V.K. Lazarov [c,d], A. S. Eggeman [a,*]

[a] Department of Materials, University of Manchester, Manchester M13 9PL, UK
[b] School of Geographical and Earth Sciences, University of Glasgow, Glasgow G12 8QQ, UK
[c] School of Physics, Engineering and Technology, University of York, York YO10 5DD, UK
[d] York-JEOL Nanocentre, University of York, York YO10 5BR, UK

ARTICLE INFO

Keywords:
Unsupervised machine learning
Correlated STEM
4D-STEM
STEM-EDS
Fuzzy clustering
Heusler alloys

ABSTRACT

In this study the effect of merging correlated energy dispersive X-ray (EDS) spectra and electron diffraction data on unsupervised machine learning (clustering) is explored. The combination of data allows second phase coherent precipitates to be identified, that could not be determined from either the individual EDS or diffraction data alone. In order to successfully combine these two distinct data types we leveraged a data fusion method where both data sets were normalised and combined using a robust scaler followed by variance equalisation. A machine learning pipeline was implemented which performs dimensional reduction with PCA and followed by fuzzy C-means clustering, as this allows signals from overlapping regions of the microstructure to be partitioned between different clusters. User control of this partition is used to confirm a change in the stoichiometry of the embedded second phase regions.

1. Introduction

There is currently a drive towards greater automation of materials characterisation in transmission electron microscopy (TEM). Faster detectors, more stable electron optics and computer control all allow for the analysis of larger areas, greater numbers of regions of interest and increased modality. One potential bottleneck in these approaches is the ability to automate data analysis, particularly the identification of features requiring additional scrutiny. Such computer vision (CV) methods can be used to reduce the burden on human analysts and to detect trends that may not be immediately obvious to even a trained user.

There are several possible approaches that can be adopted for such CV, these broadly fall into supervised and unsupervised machine learning categories. The most common supervised approach is the use of artificial neural networks for deep learning [1]. Such deep learning approaches have showed value for a range of automated data processing and data segmentation cases. One powerful example is the denoising capability, here extremely low dose annular dark-field scanning transmission electron microscopy (ADF STEM) images can be recorded, that would under normal circumstance be impossible to interpret, either through sparse scan patterns or through extremely low dose, but through deep learning analysis the lattice image can be recovered [234]. A further stage to this is training the network to recognise particular features in the image, making for automatic segmentation and hence the ability to explore a range of micro- and nanoscale features such as atomic defects [5–6] and dislocations or grain boundaries[7].

Unsupervised methods offer a slightly different approach, where instead of looking to directly label the data, these are often used to identify trends or features within the data [8 9 10] or to model the data in a more tractable form [1112]. Since there is typically no formal training of the algorithm using pre-labelled datasets, these trends arise from the statistical variance the different features introduce into the overall data, making it an attractive alternate option in STEM experiments where there are typically many more experimental measurements than unique components of the microstructure in question. This is also significant for systems where there is uncertainty about the nature of the microstructure and hence the end goal is not segmentation *per se* but the discovery of microstructural features present. This latter method will be explored in depth in this article.

Machine learning methods that focus on dimension reduction [13],

* Corresponding author.
E-mail address: alexander.eggeman@manchester.ac.uk (A.S. Eggeman).

https://doi.org/10.1016/j.commatsci.2023.112336
Received 6 March 2023; Received in revised form 4 May 2023; Accepted 14 June 2023
Available online 28 June 2023
0927-0256/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

notably principal component analysis (PCA) use explained variance as a proxy for labelled (or pre-categorized) information within a set of measurements while maintaining the global structure of the data. So experiments should be designed to maximise how category information (specific to individual microstructure elements) affects the variance present in the measurement. One approach to do this is to record multiple signals at each point in the STEM scan. Certain signals detected in the STEM can be considered as independent (or orthogonal) since the physical mechanisms producing the signals are sufficiently different, an example might be grain that exhibits a small change in orientation (bending), leading to a change in Bragg diffraction, while exhibiting unchanged EDS since the stoichiometry of the grain should be independent of orientation. Hence any covariance in these signals is not attributed to some common underlying dependence of the physical mechanisms producing the signals but it is instead an indication of related changes in the sample. Using the same example as before, a change in orientation changes the Bragg diffraction but if there is a corresponding change in EDS (stoichiometry) then this might not be the same grain, there may be an additional microstructural component (defect, inclusion, etc.) affecting the measurement.

Hence, when combined, the covariance between the different signals is expected to improve the cohesion and separation of clusters of data points, thus improving the ability of clustering algorithms to segregate features within a microstructure [14]. It is this concept and the experimental design to achieve this that will be explored in this article.

It should be noted that phase identification can be achieved manually (indeed it is currently what researchers do through intensive study) for example by inserting apertures (either physically in the microscope or virtually in the case of recorded diffraction data) to look for microstructure [1516], however this is time-intensive and is not compatible with the move to automated and high-throughput analysis. It is also possible for even experienced users to misinterpret or miss minor features that may be significant to the microstructure, lending further weight to the use of CV as a means to improve data analysis.

The test sample used was a Heusler alloy, Co_2FeSi (CFS), known to exhibit a range of coherent ordered structures (A2, B2, $L2_1$, $D0_3$ [17]). This particular system offers the challenge of completely common sublattice reflections in their diffraction patterns (common with the disordered A2 structure), which makes identification of phases from lattice imaging difficult since many of the detectable spatial frequencies in an atomic STEM image are the same in all phases. The phases also have near-identical composition making elemental mapping uninformative. Furthermore, two of the phases exhibit common superlattice reflections arising from ordering of the cobalt sublattice, which is present in both B2 and $L2_1$ (note the $L2_1$ structure has an additional ordering of the iron and silicon sublattices). As such this is the sort of material that requires a highly experienced analyst to study in detail and so represents the type of challenge that a CV approach must be able to tackle. The material is representative of the sort of advanced functional materials that high-throughput multiscale TEM characterisation would be used for.

2. Materials and methods

STEM experiments were performed on a Thermo Fisher Talos F200X TEM operated at 200 kV. During STEM scan both EDS spectra and convergent beam electron diffraction (CBED) patterns were recorded simultaneously using the STEM software trigger signal to coordinate both acquisitions. The electron beam convergence angle was 1.5 mrad, achieved using 10 μ C2 aperture. EDS was recorded using an Thermo Fisher Super-X SDD detector with 0.9 srad collection angle. Electron diffraction patterns were recorded using a Quantum Detectors Merlin Quad detector with 512x512 pixels. For the scans used in this study a dwell time of 2–25 ms was used to scan regions with a step-size of 1–2 nm. Additionally conventional ADF STEM images were recorded for each scan area.

Unsupervised clustering was performed using a probabilistic fuzzy C-means implementation [818] applied to a reduced dimensional transform of the original data using PCA. Cluster centres were iteratively updated to minimise the sum of 1st membership values for all measurements. Additionally, the Gustafsson-Kessel approach was used to allow elliptical rather than spherical clusters. [19].

The sample used was a CFS Heusler alloy thin film on a Si substrate, grown via molecular beam epitaxy (MBE). The MBE growth was achieved using two k-cell effusion evaporators for co-effusion of Fe and Co along with a Silicon Sublimation Source. TEM specimens were prepared via Focused Ion Beam (FIB) on an FEI Nova NanoLab 200 after coating with carbon and an 80:20Pt/Pd alloy. The FIB lamella, cut along the $[1\bar{1}0]$ direction of CFS, was then polished in a Gatan Precision Ion Polishing System II with a 0.1 kV argon beam.

3. Results and discussion

3.1. ADF analysis

An initial examination of the materials was performed to provide an overview of the sample structure and to determine if microstructure was readily discernable. An ADF image of the first sample is shown in Fig. 1a and for comparison a virtual ADF aperture was constructed (shown on the position averaged CBED (PACBED) pattern in Fig. 1b) resulting in the virtual ADF image of the sample shown in Fig. 1c.

From each of the STEM images the general arrangement of the different layers in the sample is evident, more importantly there is no discernable microstructure visible in the CFS layer, which is expected given the reasons outlined in the introduction. Initial scrutiny of the PACBED pattern suggests that beyond the [1 1 0] zone axis reflections from the silicon substrate and the [1 1 0] reflections of the CFS A2 sublattice there is little distinctive structure signal to focus on.

3.2. Data clustering

A significant part of the data clustering method is the correct preprocessing of the experimental data. The overall workflow is outlined in the left-hand side of Fig. 2. Individual steps in this process will be introduced in this section.

3.2.1. Dimension reduction

Data clustering first requires a reduction in the dimensionality of the problem, in the EDS measurements there are 1000 energy channels and in the CBED patterns there are 262,144 pixels, each of which is a dimension when considering the Euclidean distance (similarity) between two measurements. Not only does the high dimensionality risk the calculation of distance becoming a computational bottleneck, but it also leads to the so-called ‘curse of dimensionality’ where there is an exponential increase in volume of data space with increasing dimensions that results in all observations appearing equally sparse and dissimilar, which is detrimental to clustering that utilizes distance-based metrics [20,21]. The dimensional reduction stage transforms the data to a lower dimensional space while retaining the information in each individual measurement [7], while potentially removing unwanted noise.

For this step PCA was utilized as a linear dimension reduction technique. This technique attempts to capture the most variance with the least number of components by imposing a constraint such that every subsequent component must be uncorrelated to the prior component found. This constraint results in a decomposition that can capture the greatest variance using the least number of components. The PCA output models the data as a linear combination of the components, which are often referred to as latent variables (in this case a combined EDS spectrum and CBED pattern) that do not necessarily exist within the measurements. Each latent variable is weighted by a loading (that varies spatially) such that the weighted sum of the latent variables at each position in the scan recovers the original measurement as closely as

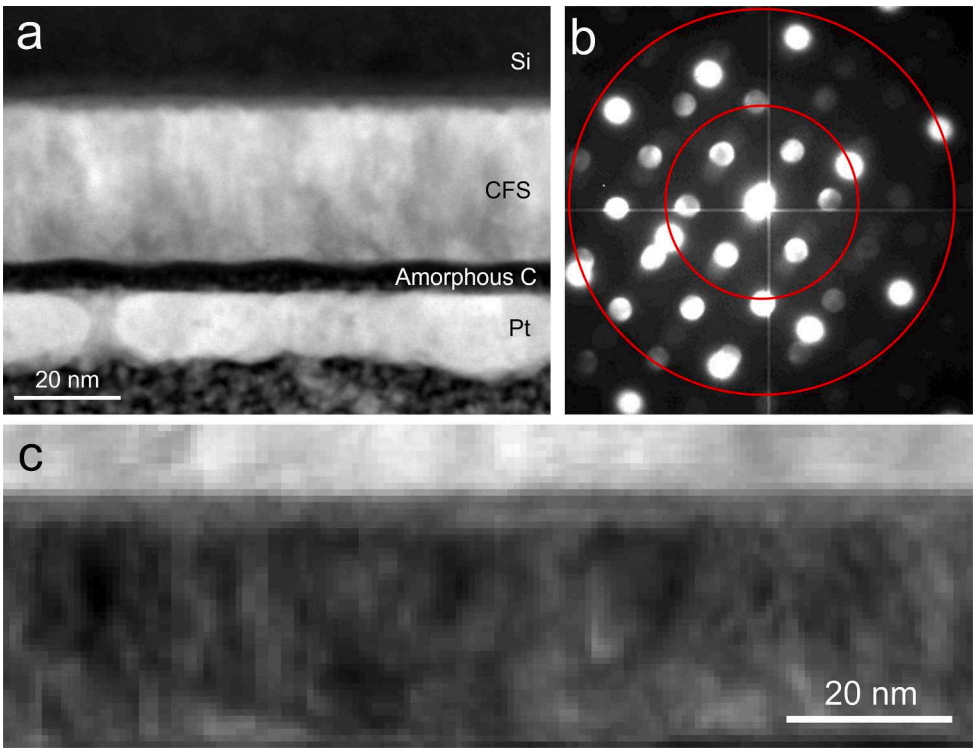


Fig. 1. A) HAADF stem image of a typical region of the cfs sample, b) PACBED pattern from a 4d-stem scan overlaid are the limit of the virtual ADF (VADF) annulus used to produce c) the VADF image of the CFS layer (lower) and the edge of the silicon substrate (upper).

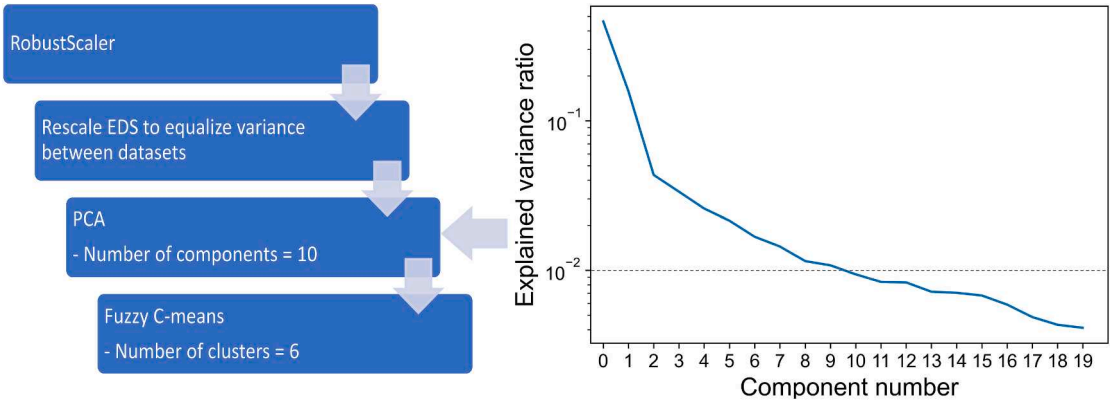


Fig. 2. Illustration of machine learning pipeline – a scree plot of explained variance was used to determine the optimal number of PCA components for the CBED data recorded from the region shown in Fig. 1.

possible.

This satisfies the dimension reduction step since each latent variable/component is now a 'dimension' and the loading values associated to each spatial point for every latent variable indicate how well a particular latent variable describes the spatial point. Hence clustering can be performed to group measurements with similar combinations of loading values.

The distribution of variance in the CBED pattern data as a function of component number (the scree plot) is shown in Fig. 2, along with a visual representation of the machine learning pipeline.

Since the clustering efficiency is adversely affected by increasing dimensionality a cut-off in the amount of information that the reduced data can describe is needed. In this study, the authors selected a cut-off of 1% in the explained variance ratio, it was felt that increasing the total explained variance (or information described by the components) by less than 1% compared to the prior components was insufficient reward for the cost of an additional dimension to be included in the clustering calculation. The line in Fig. 2 suggested that the 10 most significant factors were to be included in the clustering step in this instance. For further clustering calculations a similar result was found and the dimensionality was kept to 10 to allow consistent comparison of the machine learning outputs (see SI for more information).

3.2.2. *Signal merging*

To concatenate the two signals, the 2D CBED patterns were unfolded to form a 1D array of spatially uncorrelated pixels, hence achieving the same dimensionality with EDS spectrums. Subsequently, a pre-processing step is needed to balance the contributed variance between the two sets of measurements (step 2 of the workflow presented in Fig. 2). Essentially the two need to be scaled such that the overall variance of the combined data is not dominated by the variance present in one or the other (which would return PCA components and hence

clustering results that would be in essence identical to a single measured signal, not the combination of the two). In its raw data form, individual pixel intensities in a CBED pattern can range from zero to an order of magnitude of ten thousand, while the counts in this particular EDS dataset rarely exceed one order of magnitude. In addition, the main issue is the need to address the different structure of the data present in the two data. The EDS spectra are almost completely sparse with relative few delta-function-like emission lines. The CBED pattern is, despite first instinct, not sparse, with a significant background (thermal diffuse) over which there exist a relatively large number of significant discs of scattered intensity. The intensity in these signals therefore have significantly different distributions, making the correct choice of normalisation important. In this case, a robust scaling step which involves subtracting the median value and dividing by the interquartile range of each signal (step 1 in the workflow in Fig. 2). This was followed by the scalar multiplication of the EDS signal values by the ratio of the individual variance of each scaled dataset. This effect can be seen in Fig. 3.

This shows the output PCA components (vectorised CBED with EDS) both before and after robust scaling (Fig. 3a and 3b respectively). The information described by the PCA components is notably different, hence there will be a large change to the data space used for the subsequent clustering. In particular, the variance associated with the direct beam seems to be less prominent in the components in Fig. 3b, suggesting the components describe more of the Bragg scattering and less of the changes in and around the direct beam. Finally, a rescaling step is used to balance the total variance described by each of the CBED and EDS measurements. This was done to ensure that the components contain contributions from both the structural and chemical signals present in the data. This can be seen by the presence of spectral 'lines' appearing in the right-hand region of the PCA components in Fig. 3c, while they are absent in Fig. 3a and 3b.

The specific methodology used here is not a generally applicable solution to pre-processing for all merged data situations. There are a range of normalisation and noise correcting possibilities that can be used is very broad and specific data types and quality of recorded data will determine which approach is best. Examples of using standard scaling (subtract mean and divide by standard deviation) or Poissonian noise correction are shown in SI with differing degrees of success. The point here is that a scaling step will almost always be necessary to ensure the merged data is not dominated by one of the component parts.

3.2.3. *Clustering studies of microstructure*

The output for 10 PCA components and 6 clusters are shown in Fig. 4, in these images each measurement (pixel) is assigned a colour based on the cluster it is most strongly attributed to and the brightness of the colour is proportional to the membership value of the cluster (i.e. how close the measurement lies to the cluster centre). The clustering on the EDS data alone (Fig. 4a) identifies the silicon substrate (at the top of the image), otherwise the CFS layer appears to be a seemingly random mixture of other clusters with generally quite low association to the different clusters. This suggests a poor clustering outcome, that can be attributed to the extremely low counts in the individual spectra (given the short dwell time and the relatively low fluence with the small condenser aperture).

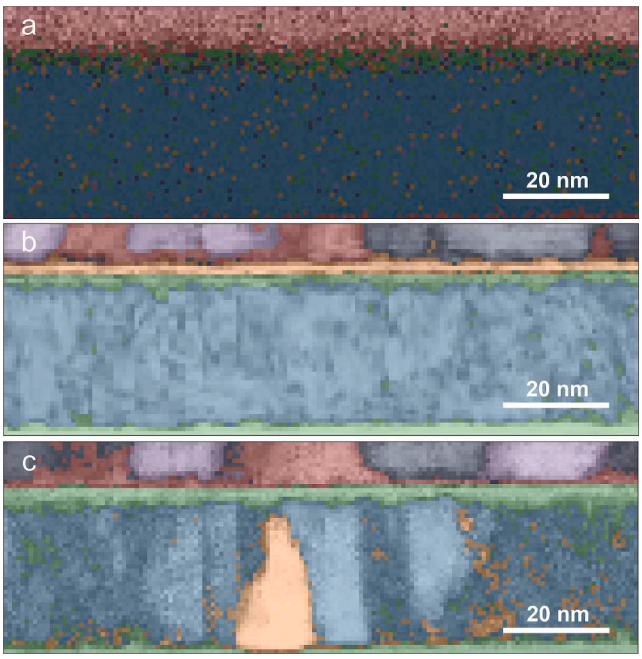


Fig. 4. Clustering results for STEM data comprising a) EDS measurements only, b) CBED measurements only and c) merged EDS and CBED measurements.

The clustering of the CBED data alone (Fig. 4b) provides slightly more insight with a range of clusters associated with the silicon substrate (likely arising from some small variations in thickness or orientation that give rise to distinctive changes in the diffraction patterns). Also, the interfaces between the CFS and the substrate/capping material are also identified, but within the bulk of the CFS little or no variation in structure is seen. The increased brightness of the colours does however suggest a reasonable partitioning of the data by the clustering process.

The significant difference arises when the two signals (EDS and CBED) are merged into a single 'measurement'. In this case the clustering result (Fig. 4c) recovers the silicon substrate variations seen in Fig. 4b but also identifies two clusters within the CFS layer, a matrix phase (blue) with a second phase present (orange) that must exhibit sufficiently distinctive diffraction and EDS signals.

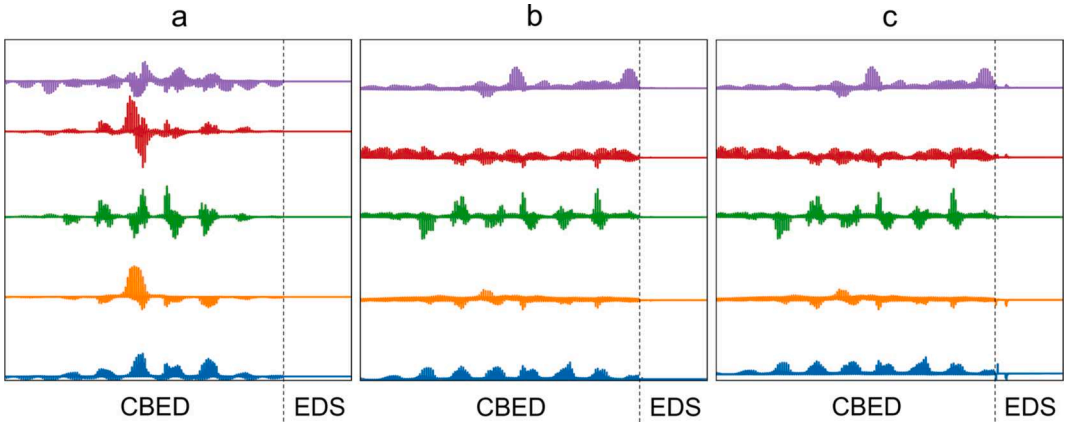


Fig. 3. The first 5 PCA components for merged EDS and CBED data, a) without robust scaling, b) with robust scaling, and c) with robust and variance scaling.

The choice of clustering algorithm was predicated on the likely nature of the data (with multiple overlapping components making a hard cluster approach unlikely to return reliable results). As with the pre-processing in 3.2.2, alternate clustering approaches were attempted on this data and their results are discussed in SI. The selection process for the number of clusters is also shown in SI.

Identification of microstructural components in this way is not limited to this material. An example of clustering analysis on a different material with similar outcomes is presented in SI.

3.3. Cluster centre analysis

One advantage of the probabilistic fuzzy clustering method used in this study is that the cluster membership values for each measurement allow a weighted mean signal for each cluster to be determined from the original experimental data. This effectively becomes an 'intelligent' guided method to highlight unique features of that cluster and suppress features attributable to other clusters (something that would be impossible with a hard clustering approach). An example is shown in Fig. 5. This shows the CFS clusters determined for data taken from a different FIB lamellae extracted from the same bulk sample as the sample shown in Fig. 4. The results in Fig. 5a-c shows the membership for the major component of the CFS layer (Fig. 5a), the cluster average diffraction (Fig. 5b) exhibits the 002 reflection (circled in red) associated with the B2 ordered structure. Alongside this the cluster average EDS signal (Fig. 5c) shows the Co Kα peak being considerably higher than the Fe Kα peak, which is expected for an overall composition of Co_2FeSi. Absolute quantification of these spectra is likely to be unreliable since there is no internal standard that can be applied to results coming out of the clustering. The fact that all measurements have a membership of all clusters leads to an inherent blending of features that makes absolute quantification open to significant errors and artefacts.

Fig. 5d-f shows the cluster attributed to the CFS second phase material for this sample, here a completely different atomic ordering is determined from the CBED pattern with superlattice reflection appearing at 1/3 ⟨442⟩* and 1/3 ⟨224⟩* type positions. For this phase the weighted EDS shows a much stronger Fe Kα peak than the Co Kα suggesting a completely different stoichiometry attributed to the different structural ordering. Given the smaller number of measurements that were attributed to the cluster shown in Fig. 5f there is an associated increase in the noise in the representative cluster centre EDS spectrum (particularly compared to the matrix cluster in Fig. 5c).

By comparison, Fig. 6a shows a second phase inclusion found in the CFS layer (albeit in a different region of the sample) where the diffraction pattern (Fig. 6a-b) has the same form as that seen in Fig. 5e but where the EDS spectrum appears similar to the bulk CFS with the Co Kα peak considerably higher than the Fe Kα. This apparently contradicts the results presented in Fig. 5d-f.

The apparent inconsistency can be explained by considering how individual measurements are attributed to clusters and how this affects the mean signal. For Fig. 5, a representative cluster centre signal is determined from all measurements that have a membership value above 0.5 for that cluster. This is the lowest threshold that defines a non-ambiguous membership (since for multiple clusters all with finite memberships there cannot be more than one cluster with a membership of 0.5 or greater). If the membership threshold is made more restrictive, then only those measurements that lie closer to the cluster centre will be included in the determination of the representative signal. This should cause the representative cluster centre results to better reflect the 'true' structure and composition since outlier measurements are not included, at the expense of potentially greater detector noise. Fig. 6 shows a comparison of the cluster memberships and representative EDS signal from a different scan region for threshold conditions of 0.5 (the minimal condition for cluster membership, in Fig. 6a and 6c) and 0.75 (a more limited condition shown in Fig. 6b and 6d).

The insets in Fig. 6a and b show the representative cluster diffraction patterns, these exhibit the same geometry as Fig. 5e indicating the same structural ordering. What becomes immediately clear is that when the threshold value for inclusion is raised, the Co Kα peak reduces and the Fe Kα peak increases in the representative EDS spectra, so the overall form of the spectrum becomes closer to the result found in Fig. 5f with a more equal ratio of Co and Fe. The cost here is that fewer measurements are included (comparing Fig. 6a and 6b) resulting in more evidence of detector noise in the representative spectra.

The implication here is that the second phase in this final sample is a nanoscale inclusion embedded within the bulk CFS matrix. The individual measurements (both EDS and CBED) involve the electron beam traversing these overlapping phases and so information about all of the different phases present is encoded into the data. By setting a higher threshold for inclusion the contribution of the measurements at the boundaries of the cluster are removed, or in other words there is a reduction in the influence of the signal from regions where there is more matrix overlap and less second phase contribution.

4. Conclusions

This article introduces a workflow for incorporating truly correlative analysis into CV for studying microstructure. In its simplest form the increased information in a merger of structural and chemical signals (CBED and EDS) improves the chance of automatic identification of microstructural features that might be missed from the analysis of either signal independently. Allowing in-depth study of secondary phases that occur in the sample.

This is made all the more remarkable given that the individual EDS spectra are almost unusable in their raw form, with insufficient SNR to make conventional analysis possible. The effect of combining this with CBED makes this seemingly useless data a powerful additional constraint on the identification of features worthy of further analysis. This raises the possibility of exploiting such covariance between the myriad combinations of signals that can be accessed in the modern TEM and SEM and in the wider field of materials characterisation.

The CV method used was chosen to reflect the complexity of overlapping coherent microstructure features in this sample, even though this approach embeds a computational cost compared to simpler hard clustering methods (as highlighted in SI). This approach was chosen because it allowed the fundamental nature of the overlapping phases in the CFS layer to be identified through the change in composition variations with cluster membership threshold. For studies with different arrangements of features other clustering approaches may be more advantageous. Density-based clustering methods (such as DBSCAN) can provide a powerful method for segmenting data, although there remains an issue around the definition of 'density' in the clustering. For the data presented here, the high degree of coherency means that even in low-dimension latent space, the measurements can appear nearly continuous in one or more dimensions and so density may not be the best means of grouping data. While it is known that certain dimension reduction approaches work better with density-based clustering algorithms (e.g. UMAP outputs are routinely used for HDBSCAN), we were not able to explore every possible combination in our workflow. Given the wide range of clustering methods available [22] there is likely to be a clustering approach suited to the needs of any individual experiment. Typical issues that might need to be considered are the number, size and uniformity of measurements in the low-dimension space, which can represent the number, size and character of different features in a microstructure.

Likewise, the choice of pre-processing steps is not universally applicable. Ultimately the success of the approach relies on the pre-processing of the data, and from this study we have found that necessary consideration must be given to dimension reduction, normalisation and scaling of the two sets of measurements before they can be merged. The specific steps taken in this study have worked in the examples used in this study, however they may not be applicable to all experiments, but

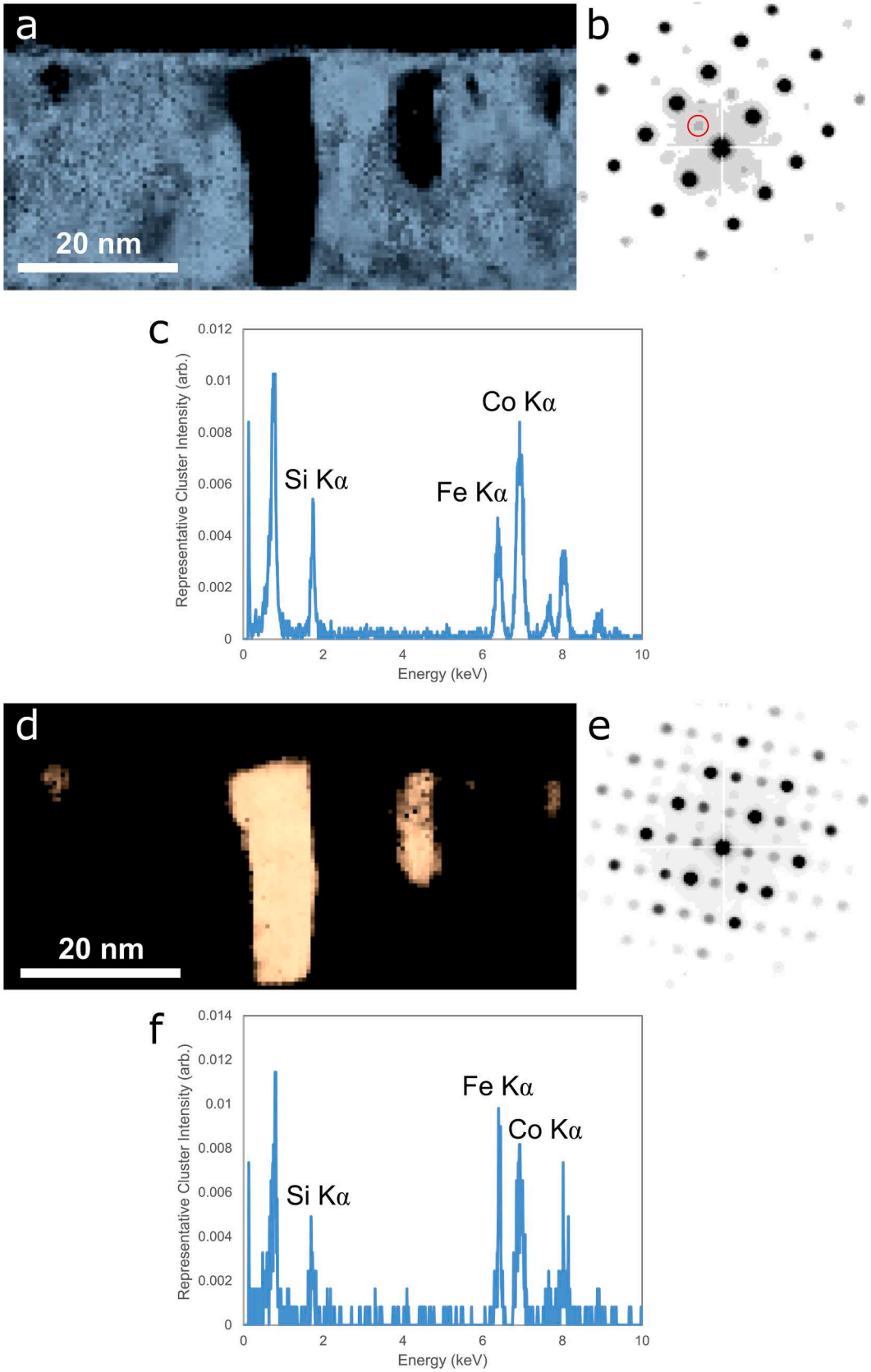


Fig. 5. Cluster analysis for a, b,c) CFS matrix and d, e, f) CFS second phases. For each cluster there is (a,d) membership map, (b,e) representative cluster centre diffraction pattern and (c,f) representative cluster centre EDS spectrum shown.

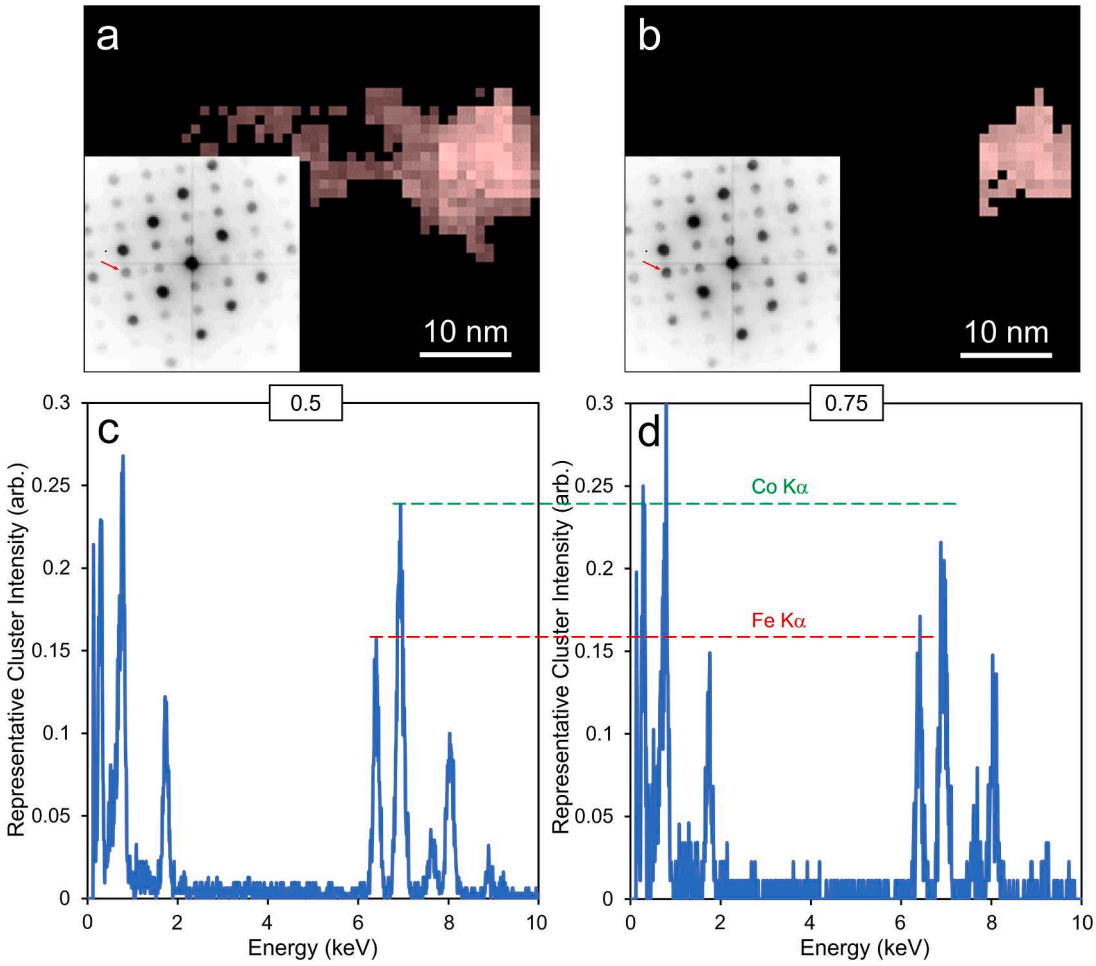


Fig. 6. Cluster analysis of a CFS second phase region showing a) and b) the cluster membership and the cluster centre diffraction pattern, c) and d) show the cluster centre EDS spectrum calculated for a threshold membership value of a,c) 0.5 and b,d) 0.75. The red arrow in the diffraction patterns highlights the change in the intensity for that reflection. The overlaid green and red lines allow comparison between the Co-Kα and Fe-Kα emission lines, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the authors believe that these are the key steps in successfully combining correlated data for such analysis.

This creates the obvious limitation that there aren't, as yet robust workflows that can be applied in a general way to any experimental data. The onus is still on the researcher to have some understanding of the specific challenges of their data and to explore the range of options for the pre-processing as well as the clustering steps, especially as there are some processing approaches that could adversely affect the analysis. Given the breadth of approaches this could be a daunting prospect, however this application is still in relative infancy and the authors envisage that, in time, these sorts of approaches will become more routine and more generally applicable. When these general approaches are realised, it is the view of the authors that multi-dimensional data such as that presented in this work will be extremely valuable as input.

CRediT authorship contribution statement

E.C. Duran: Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Z. Kho:** Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **J.F. Einsle:** Formal analysis, Methodology, Writing – review & editing. **I. Azaceta:** Investigation, Resources. **S.A. Cavill:** Funding acquisition, Supervision. **A. Kerrigan:** Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing. **V.K. Lazarov:** Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. **A.S. Eggeman:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code have been shared in the manuscript with a link to a repository.

Acknowledgements

This work was supported by the Henry Royce Institute for Advanced Materials, funded through EPSRC grants EP/R00661X/1, EP/S019367/1, EP/P025021/1, EP/S021531/1 and EP/P025498/1. ECD acknowledges financial support from the Republic of Türkiye Ministry of National Education. ASE acknowledges financial support from the Royal Society. ZK thanks the EPSRC for the studentship provided to them through the Department of Materials Doctoral Training Account. IA

acknowledges funding through EPSRC grant EP/K03278X/1.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2023.112336.

References

[1] J.M. Ede, Deep Learning in Electron Microscopy Mach, Learn. Sci. Technol. 2 (2021), 011004.
[2] J.M. Ede, R. Beanland, Partial scanning transmission electron microscopy with deep learning Sci, Rep. 10 (2020) 1–10.
[3] M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R.R. Unocic, R. Vasudevan, S. Jesse, S.V. Kalinin, Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations, ACS Nano 11 (12) (2017) 12742–12752.
[4] R. Lin, R. Zhang, C. Wang, et al., TEMImageNet training library and AtomSegNet deep-learning models for high-precision atom segmentation, localization, denoising, and deblurring of atomic-resolution images, Sci Rep 11 (2021) 5386, https://doi.org/10.1038/s41598-021-84499-w.
[5] J.C. Meyer, et al., Direct imaging of lattice atoms and topological defects in graphene membranes, Nano Lett. 8 (2008) 3582–3586.
[6] M. Ziatdinov, et al., Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations, ACS Nano 11 (2017) 12742–12752.
[7] G. Roberts, et al., Deep learning for semantic segmentation of defects in advanced stem images of steels, Sci. Rep. 9 (2019) 1–12.
[8] B. Martineau, et al., Unsupervised machine learning applied to scanning precession electron diffraction data, Adv. Struct. Chem. Imaging 5 (2019) 3.
[9] T. Bergh, et al., Nanocrystal segmentation in scanning precession electron diffraction data, J. Microsc. 279 (3) (2020) 158–167.
[10] D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791, https://doi.org/10.1038/44565.
[11] A.J. Wilkinson, Applications of multivariate statistical methods and simulation libraries to analysis of electron backscatter diffraction and transmission Kikuchi diffraction datasets, Ultramicroscopy 196 (2019) 88–98.
[12] T.P. McAuliffe, et al., Spherical-angular dark field imaging and sensitive microstructural phase clustering with unsupervised machine learning, Ultramicroscopy 219 (2020), 113132.
[13] P. Ray, S.S. Reddy, T. Banerjee, Various dimension reduction techniques for high dimensional data analysis: a review, Artif Intell Rev 54 (2021) 3473–3515, https://doi.org/10.1007/s10462-020-09928-0.
[14] C.M. Parish, Cluster Analysis of Combined EDS and EBSD Data to Solve Ambiguous Phase Identifications, Microsc. Microanal. 28 (2022) 371–382.
[15] E.F. Rauch, M. Véron, Virtual dark-field images reconstructed from electron diffraction patterns, Eur. Phys. J. Appl. Phys. 66 (2014) 10701.
[16] P. Harrison, et al., Reconstructing dual-phase nanometer scale grains within a pearlitic steel tip in 3D through 4D-scanning precession electron diffraction tomography and automated crystal orientation mapping, Ultramicroscopy 238 (2022), 113536.
[17] N. Patra, et al., Pulsed laser deposited Co_2FeSi Heusler alloy thin films: effect of different thermal growth processes, J. Alloy. Compd. 804 (2019) 470–485.
[18] B. Martineau, Scikit-cmeans's documentation. Available at: https://bm424.github.io/scikit-cmeans/ (Accessed: April 27, 2023).
[19] D. Gustafson, W. Kessel, in: Fuzzy Clustering With a Fuzzy Covariance Matrix, IEEE, San Diego, 1978, pp. 761–766.
[20] R.B. Marimont, M.B. Shapiro, Nearest neighbour searches and the curse of dimensionality, IMA J. Appl. Math. 24 (1) (1979) 59–70.
[21] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metric in high-dimensional space. In: Van den Bussche, J., Vianu, V. (eds) Database Theory — ICDT 2001 (2001).
[22] Pedregosa, et al., Scikit-learn: Machine Learning in Python, JMLR 12 (2011) 2825–2830.