Microscopy and Microanalysis, page 1 of 11
doi:10.1017/S1431927618015386

Microscopy AND Microanalysis

Original Article

The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data

Jing Wang[1], Daniel K. Schreiber[1], Nathan Bailey[2], Peter Hosemann[2] and Mychailo B. Toloczko[1]

[1]Pacific Northwest National Laboratory, Energy and Environment Directorate, Richland, WA, 99354, USA and [2]Department of Nuclear Engineering, University of California, Berkeley, CA, 94720, USA

Abstract

Atom probe tomography (APT) is a powerful technique to characterize buried three-dimensional nanostructures in a variety of materials. Accurate characterization of those nanometer-scale clusters and precipitates is of great scientific significance to understand the structure–property relationships and the microstructural evolution. The current widely used cluster analysis method, a variant of the density-based spatial clustering of applications with noise algorithm, can only accurately extract clusters of the same atomic density, neglecting several experimental realities, such as density variations within and between clusters and the nonuniformity of the atomic density in the APT reconstruction itself (e.g., crystallographic poles and other field evaporation artifacts). This clustering method relies heavily on multiple input parameters, but ideal selection of those parameters is challenging and oftentimes ambiguous. In this study, we utilize a well-known cluster analysis algorithm, called ordering points to identify the clustering structures, and an automatic cluster extraction algorithm to analyze clusters of varying atomic density in APT data. This approach requires only one free parameter, and other inputs can be estimated or bounded based on physical parameters, such as the lattice parameter and solute concentration. The effectiveness of this method is demonstrated by application to several small-scale model datasets and a real APT dataset obtained from an oxide-dispersion strengthened ferritic alloy specimen.

Key words: atom probe tomography, cluster analysis, density-based clustering, oxide-dispersion strengthened alloy

(Received 10 July 2018; revised 17 September 2018; accepted 14 October 2018)

Introduction

Atom probe tomography (APT) is a powerful technique able to characterize and visualize three-dimensional (3D) chemical information at the sub-nanometer length scale (Bas et al., 1995; Geiser et al., 2007; Geiser et al., 2009; Gault et al., 2012). It has become a standard approach to study solute clustering and fine-scaled precipitates in many materials (Miller & Kenik, 2004; Bachhav et al., 2014; He et al., 2014; Bailey et al., 2015). Accurate characterization of those nanometer-scale clusters and precipitates is of great scientific significance to understand the structure–property relationships and the microstructural evolution. For example, the size, number density and composition of oxide nanoprecipitates in oxide-dispersion strengthened (ODS) ferritic alloys are strongly correlated with the radiation resistance and mechanical properties of those materials (Certain et al., 2013; Ribis & Lozano-Perez, 2014; Chen et al., 2015; Wharry et al., 2017). Accurate and reproducible extraction of these values from 3D APT data is therefore of paramount importance to understanding and predicting the material behavior.

Several techniques have been developed to describe clustering behavior from 3D reconstructed point cloud datasets for APT. Some of them, such as isoconcentration surface (Hellman et al., 2000; Hellman & Seidman, 2002; Hellman et al., 2003), maximum separation method (MSM) (Hyde & English, 2000; Vaumousse et al., 2003; Miller & Kenik, 2004), and core-linkage (Stephenson et al., 2007) methods, have become well known. Currently, two methods, the isoconcentration surface method and a density-based cluster analysis, are most widely used because they are implemented in the commercially available software package, called interactive visualization and analysis software (IVAS, developed by CAMECA Instruments Inc., Madison, WI, USA), for analysis of APT data. The isoconcentration surface method separates clustered solute ions from those in solid solution by creating 3D surfaces of constant, user-defined concentration of selected solute species. It is easy to use, computes quickly, and is mostly successful in finding well-defined large precipitates (e.g., >5 nm diameter). However, this method suffers from drawbacks of using voxel-based algorithms which fails to identify fine-scale clusters (Hellman et al., 2003).

The density-based clustering method in IVAS is a variant of the popular density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al., 1996; Hyde & English, 2000; Vaumousse et al., 2003; Stephenson et al., 2007). The core of a cluster is defined as a group of selected data points whose k-th nearest neighbors are closer than a user-defined distance. An "envelope" technique is used to determine whether data points, other than selected ones, belong to clusters or not (Vaumousse et al., 2003) for chemical composition analysis. The current

Author for correspondence: Jing Wang, E-mail: jing.wang@pnnl.gov

Cite this article: Wang J, Schreiber DK, Bailey N, Hosemann P and Toloczko MB (2019) The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data. *Microsc Microanal.* doi: 10.1017/S1431927618015386

implementation of this method in IVAS requires three user-selected parameters (d_{max}, *Order*, and N_{min}) and an additional two parameters (envelop distance L and erosion distance $d_{erosion}$) are needed for envelope construction if chemical composition analysis is desired. Parameter selection for this method is still an ongoing challenge (Stephenson et al., 2007; Williams et al., 2013; Jägle et al., 2014; Marquis et al., 2018). The original DBSCAN algorithm requires two input parameters: the search distance ε, and the minimum number of points, *MinPts*, within search distance ε, for a given data point to be qualified as a core point of a cluster (Ester et al., 1996). Note that the cluster analysis method implemented in IVAS uses a slightly different parameter definition (d_{max}, *Order*) to qualify a core point. While d_{max} is the same as ε, *Order* is the number of neighbors of a given data point to be qualified as a core point. The difference is that *MinPts* includes the tested data point itself while *Order* does not, and thus *Order* = *MinPts* − 1. For this paper we follow the *MinPts* definition.

Cluster analysis remains an active research field in the APT community. Lefebvre et al. (2011) have demonstrated an original method based on Delaunay tessellation that requires only one input parameter, and no prior knowledge about the material is needed. During the cluster identification process, the method directly computes a sharp envelope for each cluster so that their morphology is more appropriate. Felfer et al. (2015) presented a method that computes Voronoi cells of solute ions and extracts clusters based on an automatic thresholding process which in turn is based on the cell-volume size distributions. The composition of clusters is then calculated in a way similar to that described by Lefebvre et al. (2011) using Delaunay tessellation. Zelenty et al. (2017) used Gaussian mixture models (GMM) to extract clusters on a probabilistic basis. The process requires no input parameters and is automated using the Bayesian information criterion for determining the optimal number of clusters in a dataset.

In this paper, we describe the application of a well-known cluster analysis algorithm, called ordering points to identify the clustering structures (OPTICS; Ankerst et al., 1999), and a cluster extraction algorithm we developed to detect clusters of varying atomic density in APT data. The OPTICS algorithm, which is a generalized version of DBSCAN, does not directly partition data points into clusters nor explicitly produce clustering structures, but rather creates an augmented ordering of data points. This ordered list contains clustering information that corresponds to a wide range of parameter space in DBSCAN. An automated approach is developed to extract the hierarchical cluster structures from results of the OPTICS algorithm. The new clustering method will be referred to as "OPTICS-APT" in the following text. The effectiveness of the new cluster analysis method is demonstrated on several small-scale model datasets and a real APT dataset obtained from an unirradiated ODS ferritic alloy specimen. The advantages and limitations of the OPTICS-APT method are discussed, parameter sensitivity analysis is conducted, and the clustering result is compared with the current widely used methods.

Materials and Methods

The OPTICS-APT method was applied to model artificial APT datasets for proof of concept and real APT data for testing and validation. The model datasets consist of two types of regions: regions of high density data points (corresponding to solute clusters) and regions of low density data points (corresponding to solute in the matrix). The clusters were generated by two steps: first, the centers of clusters were randomly assigned within the simulation box; data points inside each cluster were then randomly generated following Gaussian distributions along the x, y, and z axes with peaks located at the determined cluster centers and variances were set based on desired sizes. The solute atoms in the matrix was modeled by randomly adding data points to the dataset to a target concentration. A small model dataset consists of varying atomic density clusters has been generated to demonstrate the capability of the algorithm for the theory section. The average atomic density of clusters is set to the atomic density of body centered cubic pure Fe, adjusted by the APT detection efficiency of 37%.

The real APT datasets were collected from an archival MA957 specimen, a Fe–14Cr–1Ti–0.3Mo–0.25Y_2O_3 (wt%) ODS ferritic alloy, using a CAMECA LEAP 4000X HR (CAMECA Instruments Inc.). APT data were collected in laser-pulsing mode (λ = 355 nm, 60 pJ/pulse) at a base temperature of 40 K, a laser pulse frequency of 200 kHz and voltage-controlled constant detection rate of 0.3% (0.003 detected ions/pulse). The nominal detection efficiency of this tool is ~37%. All tests and validations of the OPTICS-APT method were carried out on a 64 bit Windows workstation equipped with a Hex Core 3.5 GHz Intel Xeon E5-1650 v3 CPU and 32 GB RAM.

Theory

Broadly speaking, clustering algorithms group data according to similarities based on a set of predefined or calculated criteria (Hartigan & Hartigan, 1975). Density-based clustering algorithms consider data in the same high-density region to be more similar to each other than to the remainder (i.e., low-density regions) of the dataset. The density criteria can be defined in many ways. One common measure is the number of similar data points within a certain distance from a given data point, as defined in the aforementioned DBSCAN algorithm (Ester et al., 1996). This simple criteria cannot, however, accurately extract clusters of varying atomic density. This poses a potential issue for APT data, since clusters do not always present the same atomic density in many materials due to both microstructural differences and also field evaporation artifacts (Marquis & Vurpillot, 2008). It therefore stands to reason that an improved algorithm is necessary.

Basics of the OPTICS Algorithm

The OPTICS algorithm was proposed by Ankerst et al. (1999) to overcome the intrinsic limitations of the DBSCAN algorithm to detect clusters of varying atomic densities. An accurate description and definition of the algorithmic process can be found in the original research paper. Two input parameters are required for the OPTICS algorithm: the maximum search distance ε from a data point O, and *MinPts* for the minimum number of data points within distance ε, including point O itself. The process of the OPTICS algorithm can be briefly described as follows. The algorithm starts at a random point, o, in the dataset, extracts its neighbors within a maximum search distance, ε, and puts them in a data structure called the priority queue. The corresponding reachability distance (RD) and core distance (CD) are calculated once for every point in the queue, and their rank in the priority queue is updated based on RD. The processed points are then put into an ordered list. As long as the priority queue is not empty,

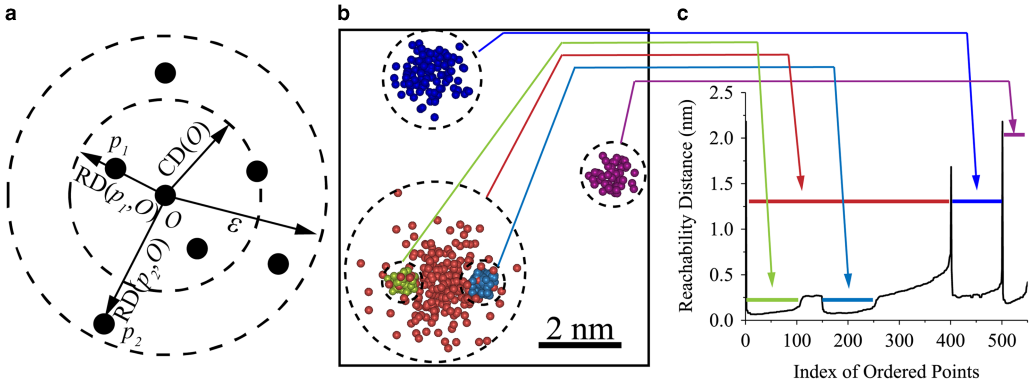

Figure 1. a: Illustration of the definitions of core distance (CD) and reachability distance (RD) in the OPTICS algorithm using *MinPts* = 4. **b:** Visualization of a small-scale model dataset consisting of five clusters highlighted by colors. **c:** Reachability distance plot (RD plot or reachability plot) of data presented in (**b**) using the OPTICS algorithm with parameters ($\varepsilon = \infty$, *MinPts* = 10).

the algorithm continues to process the data point whose RD is the smallest for the next iteration. If the priority queue is empty, the next unprocessed data point in the dataset is selected for processing. In this way, the generated ordered list guarantees the data point at index $n+1$ has the smallest RD to the data point at index n. The process is iterative and continues until all eligible points have been processed.

The definitions of CD and RD are illustrated in Figure 1a. If the number of data points within distance ε from point O (including itself) is larger than *MinPts*, point O is qualified as a core point and the CD of point O is the distance from itself to its k-th nearest neighbor, where $k = MinPts - 1$. CD is "undefined" and point o is not qualified as a core point if the total number of data points within distance ε from point O is less than *MinPts*. The *MinPts* is set to 4 for Figure 1a. RD of a point p from a core point O is either the distance between them or the CD of O, whichever is larger. RD of all points are initialized as "undefined", which is assumed to be larger than any finite value. The definition of CD and RD are based on a series of definitions and lemmas for density-based clustering, which are described in further detail in the original OPTICS and DBSCAN papers (Ester et al., 1996; Ankerst et al., 1999).

A small-scale 3D model dataset containing five clusters is shown in Figure 1b. For simplicity, the model dataset contains only clustered solute atoms. Three larger clusters in blue, red and purple were set at lower atomic densities than smaller clusters in green and cyan. The smaller clusters were intentionally embedded in red cluster to: (1) represent the challenging case of varying atomic density and (2) illustrate the cluster structure information is preserved in reachability plot produced by the OPTICS algorithm. The dataset also has practical significance, since small clusters embedded in a larger one are not uncommon in materials science research (e.g., co-precipitation or templated growth). The IVAS clustering method, based on DBSCAN, is incapable to accurately extract clusters of varying densities nor preserving information of clustering hierarchies (Ankerst, et al., 1999), while the RD plot can easily interrogate such a dataset.

The RD plot (also called a reachability plot) computed for this dataset using the OPTICS algorithm is shown in Figure 1c. The x-axis, the index of ordered points, represents the ordering of data produced by OPTICS, and the y-axis plots the corresponding RDs. These RD values can be considered as an analog measure of the local solute density in APT data, where small RD values correspond to shorter distances and thus higher local solute density. Thus atoms found in the valleys of the RD plot represent data points that are spatially close to each other (high local atomic density), and atoms in the same valley likely belong to the same cluster. The RD valleys are separated by data points of large RD, which means they are farther away from the data points in a valley. The identified clusters in the RD plot are labeled using same colors as in the visualization. The capability of the OPTICS algorithm to extract hierarchical clustering information from a single pass-through data processing is demonstrated. As illustrated by solid colored lines, both high density embedded clusters (green and cyan) and the parent cluster (red) can be clearly represented and visualized in the RD plot. Thus, it is also possible to overcome the interference of density variations within and between clusters and the nonuniformity of the atomic density in the APT reconstruction itself (e.g., crystallographic poles and other field evaporation artifacts). It is clear that the hierarchical clustering information in the RD plot can be a powerful tool for data exploration, which will be developed for APT dataset application in future.

Automatic Cluster Extraction Algorithm

Automatically extracting clustering features from an RD plot is nontrivial. The simplest way to extract the clusters is applying a global RD cutoff, $\varepsilon^{\text{cutoff}}$, below which data points with consecutive indices in the ordered list are grouped into the same clusters. According to Ankerst et al. (1999), the clustering result obtained in this way is visually indistinguishable from results obtained using DBSCAN with parameters ($\varepsilon^{\text{cutoff}}$, *MinPts*), except for some border points, with parameters satisfying $0 \le \varepsilon^{\text{cutoff}} \le \varepsilon^{\text{OPTICS}}$. Therefore, in theory, the RD plot contains clustering information equivalent to that for a wide parameter space ($\varepsilon^{\text{cutoff}}$, *MinPts*) in DBSCAN as long as $0 \le \varepsilon^{\text{cutoff}} \le \varepsilon^{\text{OPTICS}}$.

Sander et al. (2003) described an automatic method to extract clusters from an RD plot in a hierarchical scheme. It combines the strengths of density-based clustering and hierarchical clustering methods. A flowchart for cluster extraction is shown in Figure 2a. The algorithm first finds all local maxima in the RD plot, which are potential peaks to separate RD valleys, and sorts them in descending order. The algorithm then tests each local maximum to determine whether the local maximum is significant enough to divide the current node (i.e., a collection of data points from index i to index $i+j$ in RD plot) into two child nodes. The test is called the significance of separation test or simply the

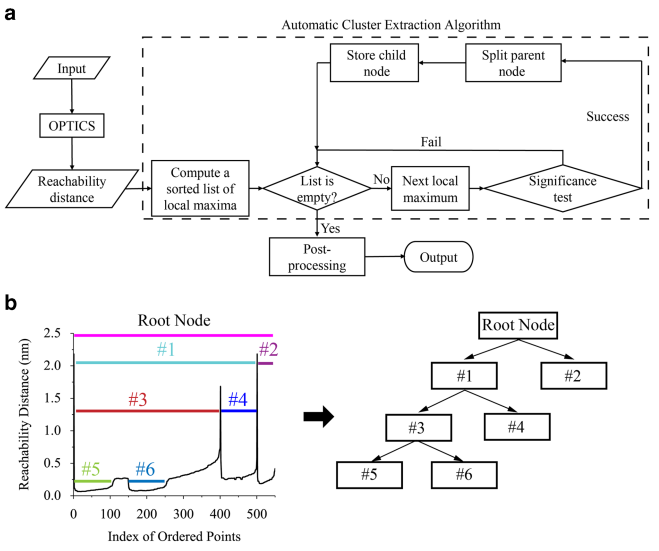

Figure 2. a: Flowchart of OPTICS-APT method. The automatic cluster extraction procedure is enclosed by dashed lines. **b:** Illustration of building hierarchy tree of nodes Node #2, #4, #5, #6 are leaf nodes.

"significance test." If the ratio between the local maximum and the average of the RDs of the potential child nodes exceeds a significance level the node is split. Taking Figure 2b as an example: the first parent node is the entire dataset, which consists of data points indexed from 0 to 550. The first local maximum is located at index 500, and two potential child nodes are indices 0–499 and 501–550. The significance test then compares the local maximum at index 500 with the average RDs of the two child nodes. Following this procedure, the root node can be divided into two child nodes #1 and #2, and node #1 can be further divided into node #3 and node #4. The value of the significance level, which sets the minimum ratio to split a parent node into two clusters, can be established between 0 and 1, but empirical experiments have shown that values of 0.7–0.8 generally give good results (Sander et al., 2003). This range matches our experience as well for human perception of APT data in our tests. The algorithm proceeds until all local maxima are exhausted, resulting in a hierarchical tree of grouped data points.

Although Sander's algorithm performs well in data with well-separated clusters, it encounters problems for realistic APT datasets. The significance test in Sander's algorithm fails when one potential child node consists of a small population of solute in clusters and a large population of solute in matrix. Because the average RD of that node is strongly biased toward the larger population of the solute in the matrix, the significance test may fail unpredictably.

To overcome this limitation and as a modification to Sander's algorithm (Sander et al., 2003), we propose to preprocess RDs in each node before the significance test. This preprocess is inspired by the observation that if a node contains both solute in clusters and solute in the matrix, the RD histogram of that node usually contains two corresponding unimodal population distributions. The simplest approach is to fit this histogram with a mixture of two unimodal distributions, and then categorize the RDs into two groups: solute in potential clusters and solute atoms in the matrix. A GMM of two components together with an expectation maximization algorithm are used during the fitting procedure for simplicity and convenience. RDs in the histogram are then assigned probabilities of being a member of either mixture component. The mean RDs for each component are calculated and the smaller mean RD, corresponding to that of a potential cluster, is used for the "significance test." Thus the preprocessing avoids the significance test being biased by the larger population of solute in matrix. The algorithm concludes when the list of local maxima is exhausted, as in the original one. Note that using the mean RD from a subset of a child node for significance test does not contradict the algorithm's capability to detect clusters of varying atomic densities, as all significance tests are performed locally within each parent node. In principle, the algorithm produces a hierarchical tree of nodes, as shown in Figure 2b, that correspond to multilevel clustering structure in which all data points, except local maxima, are preserved for further data exploration. At the end of the split procedure, a set of leaf nodes, which cannot be further divided into child nodes, are at the bottom of hierarchy. The densest local clusters are contained in these leaf nodes.

It is likely that the leaf nodes still contain a large population of solute atoms in matrix, since almost all data points are preserved when building the hierarchical tree of nodes. To separate solute in clusters from solute in matrix, a three-step postprocessing procedure is adopted. First, for each leaf node, any data with RD larger than the local maxima are discarded. Second, a GMM model is fitted to the histogram of RDs of data in each leaf node, and only those that can be categorized into a "potential cluster" component with a probability higher than 50% are kept. Third, clusters with average RD larger than a user-defined RD threshold, ρ_{th}, are filtered out to avoid extensive interference from random local density fluctuations of the solute atoms in the matrix. This process is analogous to solute atom density filtering. The RD threshold, ρ_{th}, can be estimated *a priori* in two ways, as discussed further below. Filtering clusters by size (number of points in cluster) can also be implemented as needed, but we consider it a redundancy and unnecessary for the current method.

Results

Tests on Model Datasets

The OPTICS-APT method was tested on a model dataset, which is referred as the "small-scale model dataset" in the following text, to explore its general behavior in controlled conditions. First, the effects of parameter selection on the clustering results were investigated. The maximum search distance, ε, was set to infinity so that all clustering structure information was obtained, and the number of free input parameters was reduced by 1. The minimum number of points, *MinPts*, was set to 5 and 10. The clustering results and the calculated RDs are shown in Figure 3. Figures 3a and 3d are visualizations of the detected clusters (colored spheres) and the matrix solutes (black dots). The correct number of clusters was obtained for $MinPts = 10$, while more than five clusters were detected for $MinPts = 5$. The purple and red clusters in Figure 3d were subdivided when $MinPts = 5$, as illustrated in Figure 3a. The choice of parameter may affect the number of detected clusters, similarly to other clustering algorithm such as the maximum separation method.

The reachability plots for the same analyses are presented in Figures 3b and 3e. The reachability plot is smoother for $MinPts = 10$ than 5 by reducing the influence of highly-localized density variations. Note that the exact locations of valleys and peaks are different between the two plots, because the process of ordering points is parameter sensitive in the OPTICS algorithm.

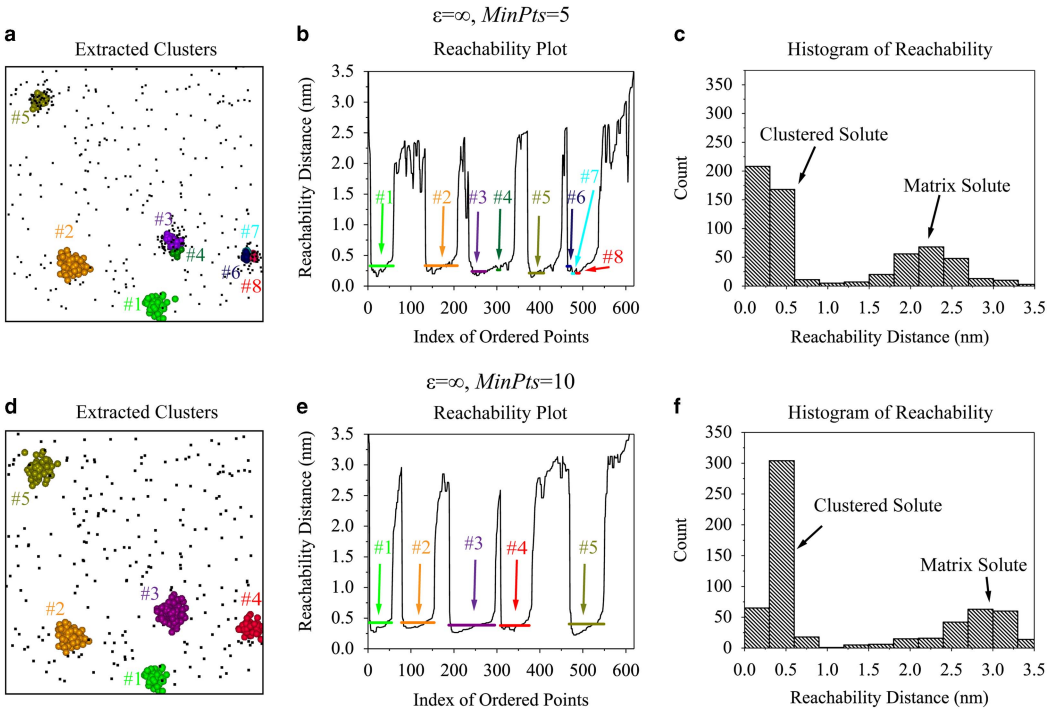

Figure 3. Effects of parameter selection for the OPTICS algorithm on the result of cluster analysis. **a,d:** Visualizations of extracted clusters (labeled with color) and the matrix solutes (black dots) from the small-scale model dataset. Note that the visualization of clustered points is enhanced using large spheres and color labels. **b,e:** Reachability plots obtained using parameters ($\varepsilon = \infty$, *MinPts* = 5) and ($\varepsilon = \infty$, *MinPts* = 10), respectively. Data points that belong to the same clusters are marked by solid lines and the colors of the solid lines are in correspondence with those in the visualization. **c,f:** Histograms of RDs, both of which follow bimodal distributions.

However, this does not affect the final clustering result, since spatially close data points are always grouped near each other. Colored solid lines are superimposed on the reachability plot to mark corresponding data points that belong to the same clusters. In Figure 3e, five dense regions of low RDs that are separated by regions of high RDs have been successfully identified. In comparison, several valleys in Figure 3b were further divided into smaller but denser regions. This is a natural result of selecting a smaller *MinPts* parameter. Since the RD, which is associated with local atomic density, is not always uniform even inside clusters, it is reasonable to encounter similar atomic density variations inside experimental clusters. Thus the automatic cluster extraction algorithm, whose goal was to extract low reachability regions based on the local environment, was merely functioning as designed. On the other hand, even if the automatic cluster extraction produces results that deviate significantly from human perception, it can be improved, with human supervision if desired, because all clustering information is available in the reachability plot. Such flexibility is one of the unique benefits provided by the RD plot versus the far more tedious sensitivity analyses performed with conventional DBSCAN methods.

The histograms of each data point's calculated RDs are shown in Figures 3c and 3f. They follow a bimodal distribution: the smaller reachability part corresponds to data points that belong to clusters, and those in the second distribution with larger reachability are usually classified as the "matrix" or nonclustered solute atoms. This plot can be used to filter out clusters that arose from random density variations of matrix solute atoms. For example, an RD of 1.0 would be a good choice for the density threshold ρ_{th} to filter out undesired clusters for both *MinPts* = 5 and 10. However, if clusters consist of only a small fraction of data, the separation becomes ambiguous. One typical case would be in dataset where the amount of matrix solute atoms is dominant.

Table 1. Comparison of the number of identified clusters and the measure of clustering accuracy (adjusted Rand index, ARI) for various cases presented previously.

	Case 1	Case 2	Case 3	Case 4
Solute atoms in clusters	388			
Solute atoms in the matrix	231	231	5231	30231
ε	∞			
MinPts	5	10	10	10
Simulated clusters	5			
Identified clusters	8	5	5	5
ARI	0.39	0.94	0.92	0.86

In the small-scale model data (Fig. 3), clusters consist of 388 data points and the matrix contains only an additional 231 solutes. In real APT datasets, it is not uncommon for only a small fraction of solute atoms to contribute to clustering, and the rest of the solute atoms are randomly distributed throughout the matrix. Thus, the robustness of the new algorithm against randomly distributed solute atoms in the matrix needs to be tested. To generate test datasets and for consistency, additional solute atoms in the matrix were added to the small-scale model dataset. A brief summary of generated datasets is listed in Table 1.

Figure 4 shows the clustering results for these test datasets. The visualizations of extracted clusters with color labels are illustrated in Figures 4a and 4b. In both datasets, the correct number of clusters have been found. Identified clusters in the dataset with 5,000 additional solute atoms in the matrix appears to be larger than those in dataset with 30,000 additional matrix solute atoms.

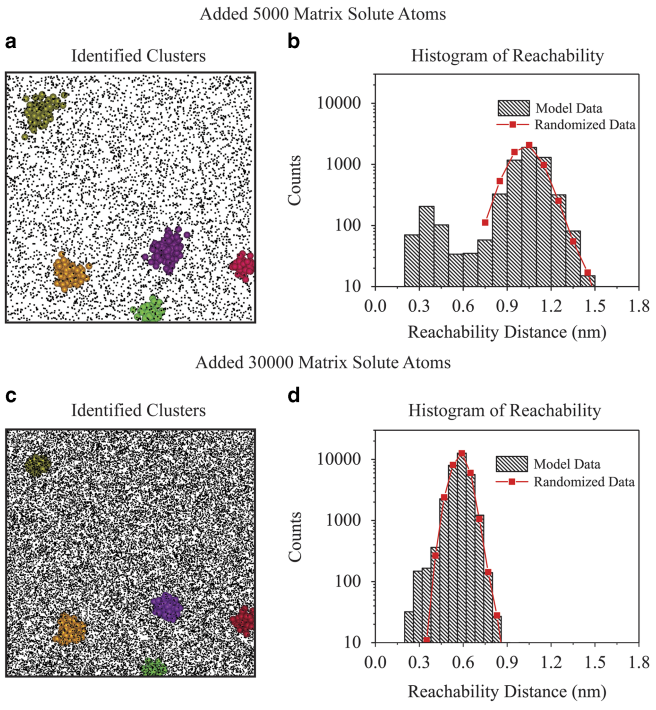

Figure 4. a,c: Visualization of identified clusters and (**b,d**) corresponding histogram of reachability for the small-scale dataset with an additional 5,000 and 30,000 matrix solute atoms, respectively, in comparison with a histogram of RD calculated for equal amounts of matrix solute with randomized positions. The input parameters were $\varepsilon = \infty$ and $MinPts = 10$. Density threshold ρ_{th} was set to 0.6 and 0.4 for 5,000 and 30,000 additional matrix solute atoms cases, respectively. The visualization of data points in clusters is enhanced using larger size spheres and color labels.

This behavior is expected, since the atomic density at the periphery of clusters becomes indistinguishable from the high concentration of solutes in the matrix, so that only core regions with higher atomic density in the cluster center are recognized. The sharpness of the cluster/matrix interface affects the clustering results (e.g., cluster sizes). At the extreme ends, the density gradient across the interface is too low and the corresponding local maximum in RD plot becomes obscure or disappears. As with MSM, in that case, the cluster is no longer separable from the matrix in the surroundings. However, the visual nature of the RD plot provides a more rapid evaluation of such extremes than in the tedious parameter optimization procedure of the MSM approach.

The histograms of reachability are plotted in Figures 4b and 4d. Due to increased matrix solute from 5,000 to 30,000, the proportion of the contribution of them to the histogram gradually became dominant, and the average RD of matrix solute decreased. Although the reachability distributions corresponding to clusters are visually difficult to detect in the histogram, a computer algorithm and statistical analysis still identifies them. For example, one method is to calculate the histogram of reachability of a randomized dataset, displayed in red lines, and apply the Kolmogorov–Smirnov test (Dodge, 2008) to determine whether two histograms are from the same distribution. If the amount of matrix solute continues to increase, eventually all clusters would become indistinguishable from the random density fluctuations of the solute atoms in the matrix.

A synopsis of these four test cases is presented in Table 1. While case 1 ($\varepsilon = \infty$, $MinPts = 5$) was heavily affected by local density variations, the other three cases successfully and correctly identified all five simulated clusters. Other than the number of identified clusters, quantitative measures for the quality of clustering can be applied. The adjusted Rand index (ARI) proposed by Hubert & Arabie (1985) is one popular measure for comparing the similarities between two sets of clustering results. The upper bound of ARI is 1, which means two clustering results are identical, and ARI is close to 0 for two randomly labeled clustering results. More details about ARI can be found in references Vinh et al. (2009) and Meilă (2003). Here, we apply ARI to quantify the accuracy of clustering results by comparing a clustering result and the nominal cluster identification of the model dataset. For case 1, the ARI is only 0.39, which is clearly less desirable than the 0.94 for case 2. With increasing matrix solutes in test datasets, the ARI decreases to 0.92 and 0.86 for cases 3 and 4, respectively. The lower ARI can be attributed to the interference from local density variation of matrix solute atoms points at higher matrix solute level. It can be seen that the OPTICS algorithm performed well in terms of accuracy of clustering for test datasets. Furthermore, the ARI results are very reasonable, suggesting it can be a useful measure to investigate parameter sensitivity and a quantitative measure to compare the effectiveness of different clustering algorithms when known class labels are available for the test data.

Tests on Real APT Dataset

In this section, the test result for the OPTICS-APT method on a full-size, real APT dataset (five million detected ions) obtained from an ODS alloy, MA957, is presented. The OPTICS-APT requires ion species selection by users for cluster identification, just like the maximum separation distance implemented in the IVAS. Typical oxide particles are composed of the elements Y, Ti, and O. Our practical experiences with this specific alloy and previous studies suggest that the most representative ionic species for oxide particles and cluster identification are $TiO^{1+,2+}$, $Y^{2+,3+}$, and YO^{2+}. An ion map of those three ionic species is shown in Figure 5a. Most of the targeted ionic species are strongly partitioned to clusters and the matrix concentration is very low (<0.2 ionic%). This dataset was selected because it contains many clusters with varying atomic densities and sizes, as illustrated in Figure 5b. There are two categories of clusters coexisting in the dataset: one group of tight clusters with a high-density core of solute ions, colored in red, and the other group of loose clusters with less-dense solute ions in cluster cores. This is a typical case

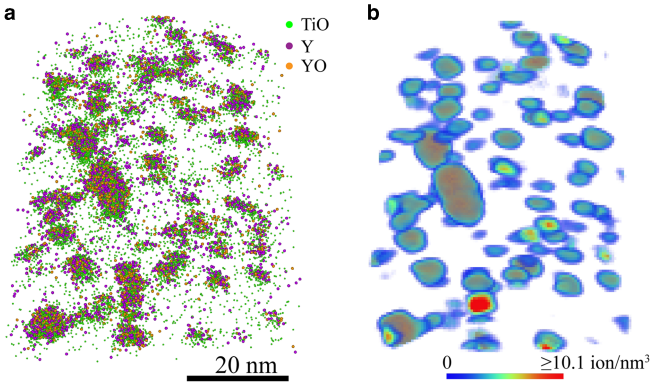

Figure 5. a: TiO, Y, and YO ion map in a real APT dataset obtained from the ODS ferritic alloy MA957; (**b**) a 3D volumetric rendering map for solute ions. The red color indicates high TiO, Y, and YO ion density while the blue indicates low density. The voxel size used to generate (**b**) is 1.0×1.0×1.0 nm and the delocalization setting is 3.0×3.0×1.5 nm.

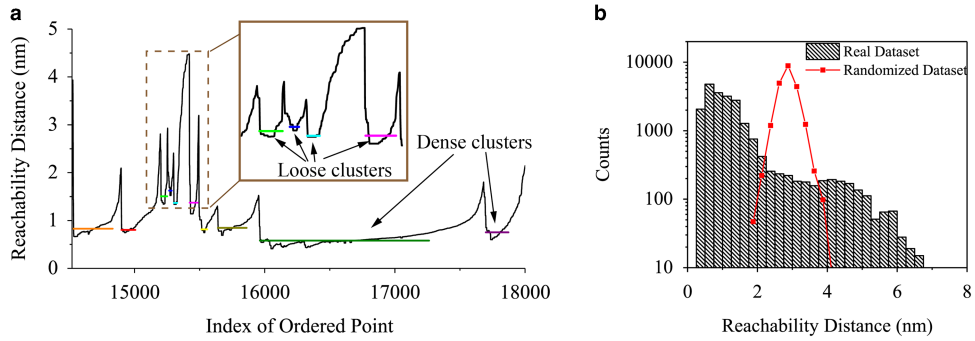

Figure 6. a: A portion of the reachability plot calculated for the real APT dataset using *MinPts* = 20. The colored solid lines indicate the range of data points belonging to detected clusters of the same color. It can be seen that the new algorithm successfully detected clusters of varying atomic densities. The start and end of the solid horizontal lines indicate the exact start and end indices of a cluster. The color of the lines does not correspond to the color of clusters. **b:** Comparison between RD histograms obtained for the real APT dataset and the corresponding randomized dataset.

where a single density threshold, like DBSCAN, would have difficulty in accurately extracting clusters, but is what OPTICS has been designed to solve.

Figure 6a shows part of the reachability plot for the real APT dataset. The start and end points of colored solid lines mark the range of data points in clusters of the same color. The coordinates of these lines along the *y*-axis were determined by the average RDs of the corresponding clusters. Clusters with both low and relatively high RDs were successfully identified using our automatic cluster extraction algorithm. Because the reachability and atomic density exhibit a quantitatively inverse relationship, this plot shows that the new clustering algorithm can identify and extract clusters of varying reconstructed atomic densities. The reachability plot clearly shows that if a single RD threshold (equivalent to a single density threshold) is used to extract clusters, not all locally dense regions can be resolved properly. The reachability plot thus provides a visual means for human supervision to further optimize the clustering results in an efficient manner. The histograms of RDs for the real APT dataset and the randomized dataset are compared in Figure 6b. Due to a significant portion of data points belonging to clusters in the real APT dataset, the effective concentration of solutes in the matrix is lower than that in the randomized dataset, which leads to larger average RDs than the randomized datasets.

The visualizations of identified clusters in the real APT data are shown in Figures 7a–7e for *MinPts* = 5–50. ε is fixed at 10 nm for all three tests to balance the completeness of the search and computation efficiency. Here completeness means that all eligible data points receive a calculated RD value. The trend of cluster visualization with respect to input parameter is very similar to results from the small-scale model dataset: when *MinPts* is small, the random local density variations inside clusters subdivide large clusters into many smaller ones. This effect is reduced by increasing *MinPts*. The algorithm also correctly identifies several parent clusters using *MinPts* of 15, as indicated by arrows. The parameter *MinPts* also sets the lower bound of detectable cluster sizes (i.e., no clusters with data points fewer than the value of *MinPts* can be identified). Thus, there is always a trade-off in parameter selection. Figures 7f–7j shows cluster size distributions with varying *MinPts* values. The size was estimated assuming precipitates are spheroidal using $R = \sqrt{\frac{5}{3}}R_g$, where R_g is the radius of gyration. The major differences among size distributions lie in the population of ultrafine clusters. As expected, and consistent with visualization, a significant portion of clusters smaller than 1 nm were detected using parameter *MinPts* = 5. At higher *MinPts* values, the percentage of ultrafine clusters decreases and the average cluster size increases. On the other hand, size distributions with peaks near 2 nm are quite similar among all three test cases in terms of absolute quantity. The number of detected clusters decreases with increasing *MinPts* values.

Besides the number density and size distribution, the composition of clusters is another feature of interest in these APT data. A full composition analysis algorithm has not been implemented yet, and it is outside the scope of this investigation. However, it is useful to generate ratios among ions that are selected for cluster analysis. Ternary plots for the three selected ion species are shown in Figures 7k–7o. The data exhibit significant scatter for *MinPts* = 5, while they are more self-consistent for *MinPts* = 10–50. This behavior can be explained by the difference in cluster size combined with detection efficiency in APT: when cluster size is small, i.e., only a few spatially close data points, the chance of that cluster having statistically representative composition is low. The corresponding counting error can also be estimated. For example, assuming a binary cluster AB of 1:1 stoichiometry (50% A), the probabilities of the detected clusters having a composition of A between 40 and 60% are 0.73, 0.76, 0.83, and 0.95 for a cluster size of 5, 10, 20, and 50 ions, respectively.

Discussion

Advantages and Limitations of the Proposed Method

The new clustering method described is designed to overcome several drawbacks of the current cluster analysis algorithm in IVAS, to create a reliable, reproducible, and easy to use cluster analysis method for APT data. One important advantage over the DBSCAN-like methods is the native capability to detect clusters of varying atomic densities in a single execution. The clustering process in the conventional DBSCAN method is equivalent to extracting clusters with a single global threshold density. This global approach neglects several experimental realities, such as density variations within and between clusters and the non-uniformity of the atomic density in the APT reconstruction itself (e.g., crystallographic poles and other field evaporation artifacts). Although the OPTICS algorithm utilizes similar input parameters (ε, *MinPts*) as DBSCAN, there is more flexibility in their implementation via RD plots to extract clusters based on their estimated local atomic density to account for these relevant variabilities.

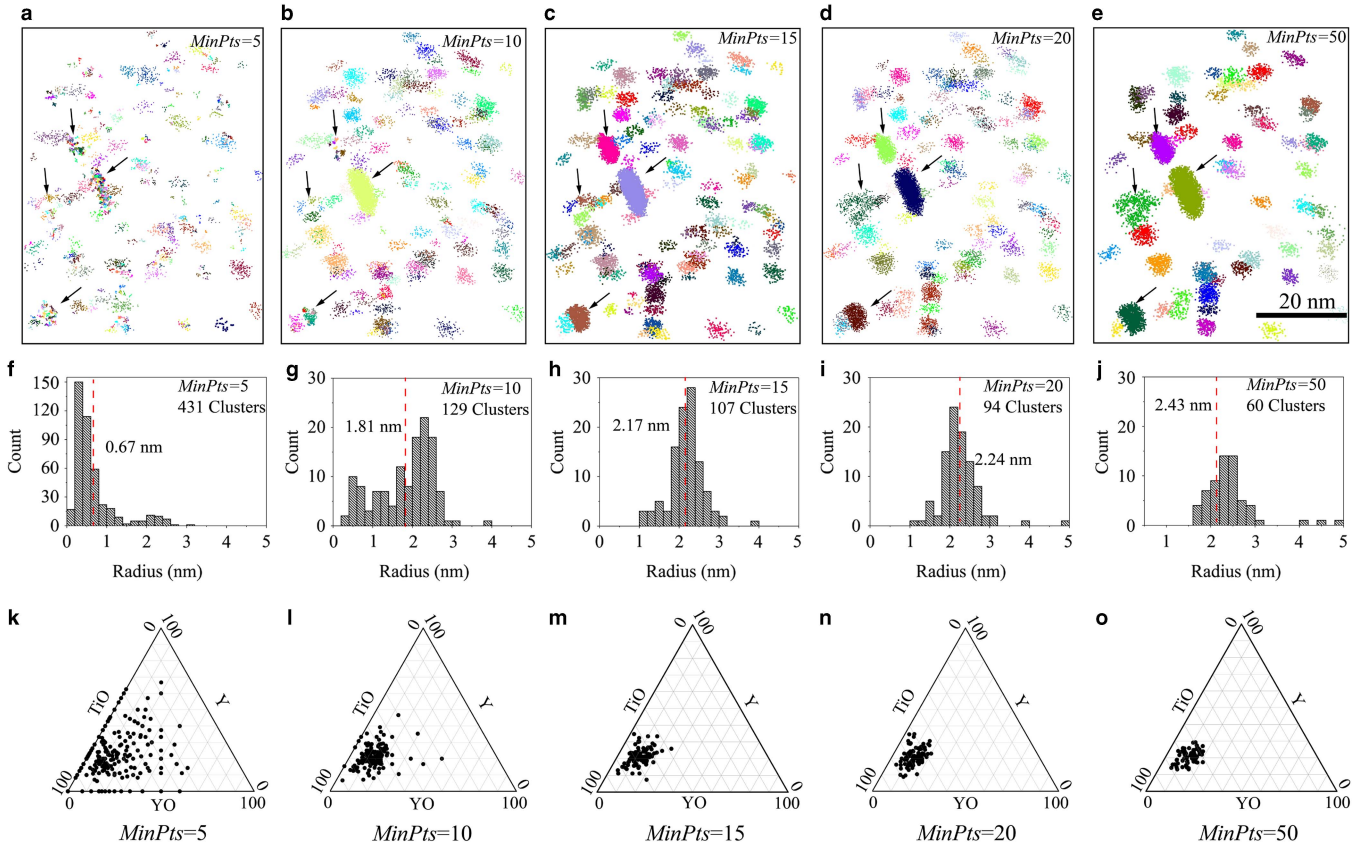

Figure 7. a–c: visualizations of identified clusters, the arrows point to several representative clustering results that evolve as parameters change; (**d–f**) corresponding size distributions of identified clusters. Red dashed lines indicate average values of the size distributions. ε is set at 10 nm for all conditions. Particle sizes were calculated by converting from Guinier radius assuming a spherical geometry.

OPTICS-APT also offers a simpler parameter selection and cluster analysis process. Because one of its two parameters, ε, can be effectively set to infinity to cover all possible distances, the only remaining free parameter is *MinPts*. The obtained RD plot is also visually intuitive for cluster identification and thus refinement of cluster extraction parameters. Since the calculated RDs for each data points are analogously inverse to the atomic density in the local surroundings of that data point, a natural way to extract clusters is to select "valley" regions in the RD plot. This process can be achieved automatically or manually. The automatic cluster extraction process demonstrated in this paper requires one user input, a density threshold ρ_{th}, to reduce the interference from density fluctuations of solute in the matrix. This density threshold can be estimated by fitting the RD histogram with two Gaussian distributions and finding the RD value where the matrix become dominant. The significance level during the significance test can potentially change based on user's needs, but in our experience, this value rarely needs to be changed as human perception matches the general output in many tests (Sander et al., 2003). It is also worth noting that OPTICS-based cluster analysis is backward compatible with DBSCAN, where the clustering result can be indistinguishable from DBSCAN if clusters are extracted from the RD plot by using a single global density cutoff (Ankerst et al., 1999).

The computation time of the OPTICS-APT method is reasonable for most APT applications on a stand-alone computer workstation. The OPTICS algorithm alone can achieve a time complexity of $O(n\log n)$ (Ankerst et al., 1999), where n is the number of data points. Similar scaling exists for the IVAS implementation of DBSCAN. The fitting of GMM via expectation maximization during cluster extraction generally has a higher time complexity and depends on several factors. Our tests show that the current implementation of the whole cluster extraction process has a time complexity of roughly $O(n^{1.37})$. For example, the cluster analysis on the small-scale model data, with about 600 points, requires ~5 s to finish, and it took <3 min (including reading and writing data to disk) for selected ions of ~21,000 solute data points in the real APT dataset presented in the results. The performance can be further improved if parallel implementations of OPTICS can be added (Patwary et al., 2013; Deng et al., 2015).

No algorithm can provide a universal solution to all problems. The major limitation of the new clustering method is that it expects a certain level of density gradient at the cluster/matrix interface in the data and defines clusters based on density differences. The algorithm does not work well if the density gradient between cluster and matrix is so low that local maximum in RD plot is not significant enough to distinguish them. In this sense, high solute concentration (such as detecting NiMnSi clusters in austenitic steels, or α' in 14Cr ferritic steels) is not a fundamental limiting factor for cluster identification; it is the differences or local density contrast between clusters and the matrix that governs the detection limit. If two clusters are somehow connected, i.e., necking (Mao et al., 2007), the algorithm may have difficulty separating them unless the necking regions have a significantly lower atomic density of solutes than in the clusters. For cases with a lot of necking between clusters, a recently reported method that uses random projection may perform better (Maurus & Plant, 2016).

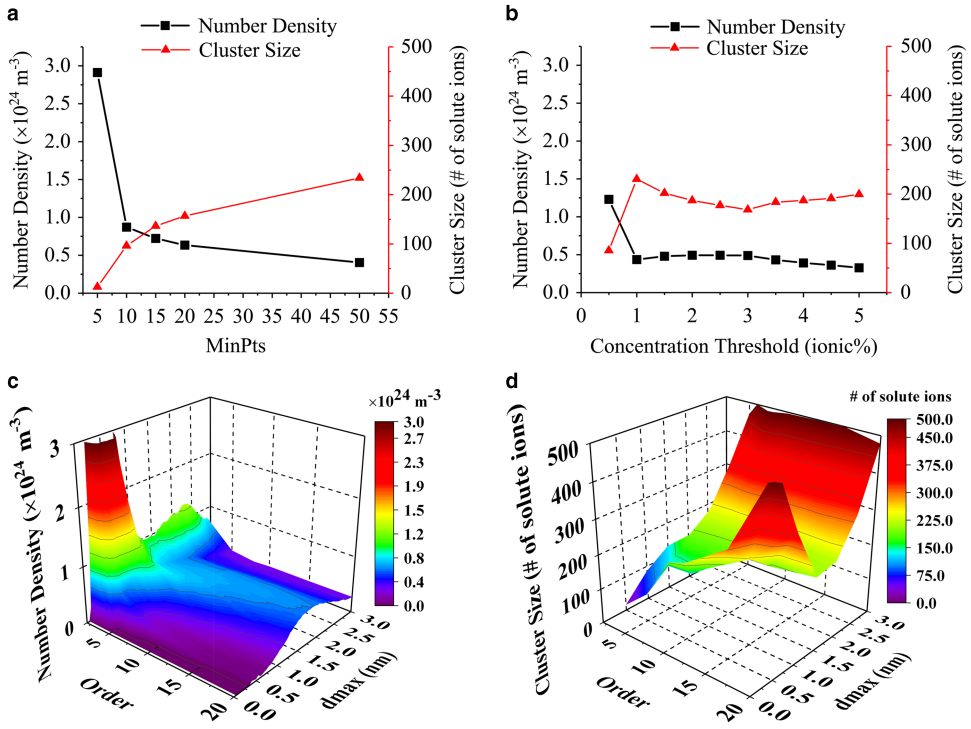

Figure 8. Sensitivity analysis of different clustering algorithm: (**a**) the OPTICS-APT method; (**b**) isoconcentration surface method; (**c**) and (**d**) parameter optimization for cluster analysis within IVAS. The voxel size used to generate (**b**) is 1.0×1.0×1.0 nm and the delocalization setting is 3.0×3.0×1.5 nm.

Sensitivity Analysis

Understanding the sensitivity of clustering results to input parameters is important to obtain accurate and reliable measurements. The real APT dataset presented in Figure 5 was used for further sensitivity analyses. In this section, the size of clusters was measured by the number of ions in order for comparison across different clustering methods. The effects of parameter selection on the number density and size distribution of identified clusters for the OPTICS-APT, isoconcentration surface method, and IVAS clustering method are presented in Figure 8. The parameter spaces for measured quantities using three methods all contain regions where measurable quantities are relatively insensitive to parameter change (sometimes called stable regions). For the OPTICS-APT method, this stable parameter space represents a balance between discarding fine-scale clusters and incorrectly subdividing larger particles into smaller ones. The trend is also expected for the isoconcentration surface method. When the isoconcentration value is higher than the upper bound of the random fluctuations in the matrix and lower than the peak concentration of clusters, a similar number of clusters can be detected. At the same time, increasing the concentration threshold value shifts the surface toward higher concentration regions, decreasing the apparent diameter of the clusters. Note that the isoconcentration surface method is successful due to the nature of this particular test dataset, which consists of many well-defined large clusters. However, the isoconcentration surface method has serious and well-known issues when detecting fine clusters or when a high solute content exists in the matrix (Marquis & Hyde, 2010).

While sensitivity analysis on the cluster analysis method in IVAS is only moderately computationally expensive, it is overtly tedious and oftentimes ambiguous for the three free parameters (*Order*, d_{max} and N_{min}). For such a DBSCAN variant algorithm, the parameter N_{min}, the minimum number of points in a cluster,

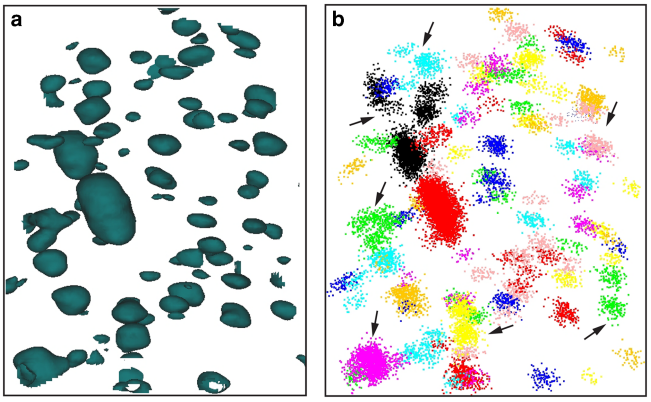

Figure 9. Clustering results using (**a**) isoconcentration surface method with the threshold set to 3 ionic%. (**b**) DBSCAN-like method in IVAS with $d_{max} = 2.0$, N_{min} = order = 15. Note that arrows in (**b**) indicate falsely connected clusters (same color).

is essentially a postprocessing redundant parameter rather than an input parameter that directly affects density-based clustering processes. For simplicity, N_{min} is set to equal *Order* here in IVAS for all tests. The number density and average size of detected clusters with respect to input parameters are presented in Figures 8c and 8d, respectively. Great variations of cluster number density can be observed in the explored parameter space. A plateau can be seen in regions $d_{max} = 1.5–2$ nm and *Order* = 10–20.

Although the clustering result across the three methods has exhibited good agreement in the current parameter space, it does not necessarily mean that the clustering results are equally accurate. Figure 9 shows an example of the clustering results using isoconcentration surface and IVAS clustering method with parameters obtained in the "stable" parameter space. Combined with the ion map in Figure 5 and the OPTICS-APT clustering

result in Figure 7c, a quick visual inspection can be performed. It is clear that the OPTICS-APT result using *MinPts* = 15 and iso-concentration surface using threshold = 3 ionic% are visually similar to the ion maps. However, many of the identified clusters in Figure 9b are falsely connected (indicated by arrows) even though simple visual inspection and other clustering method can distinguish them. We are not sure what caused this yet, as the IVAS code is proprietary. Through this simple demonstration, we see that simply choosing "stable" or "parameter insensitive" variables is insufficient to achieve accurate quantification of the clusters and results should be visually inspected for consistency.

Future Improvements

The current computer program only implemented the OPTICS algorithm and the automatic cluster extraction algorithm. Many improvements can be envisioned. For example, an interactive graphical user interface that allows users to interact with reachability plots, setting local density thresholds, and displaying corresponding changes in the visualization of an APT dataset, would greatly increase the flexibility and approachability of the method. Similarly, easier methods to explore the hierarchical clustering structure are also needed. These interface improvements, in addition to batch processing of multiple APT datasets, would be of great practical importance in realizing the potential of this analysis method. The capability of chemical analysis can also be introduced in the future by following the same envelop and erosion steps used in the current IVAS cluster analysis. Additional quantitative abilities, including schemes for evaluation of clustering quality, such as ARI, would further extend the analytical value of this approach.

Summary

In this study, a new cluster search method for APT data, OPTICS-APT, was proposed and demonstrated. It overcomes the theoretical limitations of the conventional DBSCAN-like method in the IVAS software by utilizing the more advanced and generalized clustering algorithm the OPTICS. The effects of parameter selection and robustness to a high concentration matrix solute atoms were tested using several small-scale model datasets. The effectiveness of the new clustering method to discover clusters of various densities was demonstrated in a real APT dataset obtained from the ODS alloy MA957. The new method introduces reachability plots from the OPTICS algorithm and it can be analogously used as a measure of the local density of data points. Thus, the reachability plot adds some visual clarity to parameter selection and enables the versatility for handling clusters of many different densities simultaneously. Conceptually, the new method can be programmed to be interactive for cluster extraction once a reachability plot has been computed. This would allow users to fully exploit the hierarchical nature of this method for exploratory clustering. It is worth noting that all test model datasets and the real APT datasets were focused on well-defined, density separable clusters. Some types of issues that are commonly considered challenging, such as ultra-fine scale clusters with <5 solute ions, and obscure cluster/matrix boundary, are not addressed here.

Acknowledgments. The authors would like to thank Karen Kruska at Pacific Northwest National Laboratory for providing much useful feedback during the development. This research was funded by the Fuel Cycle R&D Program Core Materials research area sponsored by the US Department of Energy, Office of Nuclear Energy. Pacific Northwest National Laboratory is operated for the US Department of Energy by Battelle Memorial Institute under contract DE-AC05-76RL01830. A portion of the research was performed using EMSL (proposal number 49117 and 49582), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research.

References

Ankerst M, Breunig MM, Kriegel H-P and Sander J (1999) OPTICS: Ordering points to identify the clustering structure, Davidson S, Faloutsos C (Eds) In *ACM Sigmod Record*, pp. 49–60. New York, NY: ACM.

Bachhav M, Robert Odette G and Marquis EA (2014) α′ Precipitation in neutron-irradiated Fe–Cr alloys. *Scripta Mater* **74**, 48–51.

Bailey NA, Stergar E, Toloczko M and Hosemann P (2015) Atom probe tomography analysis of high dose MA957 at selected irradiation temperatures. *J Nucl Mater* **459**, 225–234.

Bas P, Bostel A, Deconihout B and Blavette D (1995) A general protocol for the reconstruction of 3D atom probe data. *Appl Surf Sci* **87**, 298–304.

Certain A, Kuchibhatla S, Shutthanandan V, Hoelzer D and Allen T (2013) Radiation stability of nanoclusters in nano-structured oxide dispersion strengthened (ODS) steels. *J Nucl Mater* **434**(1), 311–321.

Chen T, Aydogan E, Gigax JG, Chen D, Wang J, Wang X, Ukai S, Garner F and Shao L (2015) Microstructural changes and void swelling of a 12Cr ODS ferritic-martensitic alloy after high-dpa self-ion irradiation. *J Nucl Mater* **467**, 42–49.

Deng Z, Hu Y, Zhu M, Huang X and Du B (2015) A scalable and fast OPTICS for clustering trajectory big data. *Cluster Comput* **18**(2), 549–562.

Dodge Y (2008) *Kolmogorov–Smirnov Test*. New York, NY: Springer New York.

Ester M, Kriegel H-P, Sander J and Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, Simoudis E, Han J and Fayyad U (Eds). In *Kdd*, pp. 226–231. Palo Alto, CA: AAAI Press.

Felfer P, Ceguerra A, Ringer S and Cairney J (2015) Detecting and extracting clusters in atom probe data: A simple, automated method using Voronoi cells. *Ultramicroscopy* **150**, 30–36.

Gault B, Moody MP, Cairney JM and Ringer SP (2012) *Atom Probe Microscopy*. New York, NY: Springer Science & Business Media.

Geiser B, Larson D, Oltman E, Gerstl S, Reinhard D, Kelly T and Prosa T (2009) Wide-field-of-view atom probe reconstruction. *Microsc Microanal* **15**(S2), 292–293.

Geiser BP, Kelly TF, Larson DJ, Schneir J and Roberts JP (2007) Spatial distribution maps for atom probe tomography. *Microsc Microanal* **13**(6), 437–447.

Hartigan JA and Hartigan J (1975) *Clustering Algorithms*. New York, NY: Wiley.

He J, Wan F, Sridharan K, Allen TR, Certain A, Shutthanandan V and Wu Y (2014) Stability of nanoclusters in 14YWT oxide dispersion strengthened steel under heavy ion-irradiation by atom probe tomography. *J Nucl Mater* **455**(1), 41–45.

Hellman OC, du Rivage JB and Seidman DN (2003) Efficient sampling for three-dimensional atom probe microscopy data. *Ultramicroscopy* **95**, 199–205.

Hellman OC and Seidman DN (2002) Measurement of the Gibbsian interfacial excess of solute at an interface of arbitrary geometry using three-dimensional atom probe microscopy. *Mater Sci Eng A* **327**(1), 24–28.

Hellman OC, Vandenbroucke JA, Rüsing J, Isheim D and Seidman DN (2000) Analysis of three-dimensional atom-probe data by the proximity histogram. *Microsc Microanal* **6**(05), 437–444.

Hubert L and Arabie P (1985) Comparing partitions. *J Classification* **2**(1), 193–218.

Hyde JM and English CA (2000) An analysis of the structure of irradiation induced Cu-enriched clusters in low and high nickel welds. *MRS Online Proc Library Arch* **650**, R6.6. doi: 0.1557/PROC-650-R6.6.

Jägle EA, Choi P-P and Raabe D (2014) The maximum separation cluster analysis algorithm for atom-probe tomography: Parameter determination and accuracy. *Microsc Microanal* **20**(06), 1662–1671.

Lefebvre W, Philippe T and Vurpillot F (2011) Application of Delaunay tessellation for the characterization of solute-rich clusters in atom probe tomography. *Ultramicroscopy* **111**(3), 200–206.

Mao Z, **Sudbrack CK**, **Yoon KE**, **Martin G and Seidman DN** (2007) The mechanism of morphogenesis in a phase-separating concentrated multicomponent alloy. *Nat Mater* **6**(3), 210–216.

Marquis EA, **Araullo-Peters V**, **Dong Y**, **Etienne A**, **Fedotova S**, **Fujii K**, **Fukuya K**, **Kuleshova E**, **Lopez A**, **London A**, **Lozano-Perez S**, **Nagai Y**, **Nishida K**, **Radiguet B**, **Schreiber D**, **Soneda N**, **Thuvander M**, **Toyama T**, **Sefta F and Chou P** (2018) On the use of density-based algorithms for the analysis of solute clustering in atom probe tomography data. In *Proceedings of the 18th International Conference on Environmental Degradation of Materials in Nuclear Power Systems – Water Reactors: Volume 2*, Jackson JH, Paraventi D and Wright M (Eds), August 13–17, 2017, pp. 881–897, Portland, OR: Cham: Springer International Publishing.

Marquis EA and Hyde JM (2010) Applications of atom-probe tomography to the characterisation of solute behaviours. *Mater Sci Eng R Rep* **69**(4), 37–62.

Marquis EA and Vurpillot F (2008) Chromatic aberrations in the field evaporation behavior of small precipitates. *Microsc Microanal* **14**(06), 561–570.

Maurus S and Plant C (2016) Skinny-dip: Clustering in a sea of noise, Krishnapuram B and Shah M (Eds), In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13–17, 2016, pp. 1055–1064, San Francisco, CA: ACM.

Meilă M (2003) Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, COLT/Kernel 2003, Washington, DC, USA, August 24–27, 2003, Schölkopf B and Warmuth MK (Eds), pp. 173–187. Berlin, Heidelberg: Springer Berlin Heidelberg.

Miller M and Kenik E (2004) Atom probe tomography: A technique for nanoscale characterization. *Microsc Microanal* **10**(03), 336–341.

Patwary A, **Mostofa M**, **Palsetia D**, **Agrawal A**, **Liao W-K**, **Manne F and Choudhary A** (2013) Scalable parallel optics data clustering using graph algorithmic techniques, Gropp W (Ed.), In *SC '13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 17–22 Nov, 2013, pp. 1–12, Denver, CO. New York, NY: ACM.

Ribis J and Lozano-Perez S (2014) Nano-cluster stability following neutron irradiation in MA957 oxide dispersion strengthened material. *J Nucl Mater* **444**(1), 314–322.

Sander J, **Qin X**, **Lu Z**, **Niu N and Kovarsky A** (2003) Automatic extraction of clusters from hierarchical clustering representations, Wang K, Jeon J, Shim K, Srivastava J (Eds), In *Advances in Knowledge Discovery and Data Mining*, pp. 75–87. Berilin, Heidelberg: Springer.

Stephenson LT, **Moody MP**, **Liddicoat PV and Ringer SP** (2007) New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data. *Microsc Microanal* **13**(06), 448–463.

Vaumousse D, **Cerezo A and Warren P** (2003) A procedure for quantification of precipitate microstructures from three-dimensional atom probe data. *Ultramicroscopy* **95**, 215–221.

Vinh NX, **Epps J and Bailey J** (2009) Information theoretic measures for clusterings comparison: Is a correction for chance necessary?, Danyluk A (Ed.), In *Proceedings of the 26th Annual International Conference on Machine Learning*, June 14–18, 2009, pp. 1073–1080, Montreal, QC, Canada. New York, NY: ACM.

Wharry JP, **Swenson MJ and Yano KH** (2017) A review of the irradiation evolution of dispersed oxide nanoparticles in the b.c.c. Fe–Cr system: Current understanding and future directions. *J Nucl Mater* **486**, 11–20.

Williams CA, **Haley D**, **Marquis EA**, **Smith GD and Moody MP** (2013) Defining clusters in APT reconstructions of ODS steels. *Ultramicroscopy* **132**, 271–278.

Zelenty J, **Dahl A**, **Hyde J**, **Smith GD and Moody MP** (2017) Detecting clusters in atom probe data with Gaussian mixture models. *Microsc Microanal* **23**(2), 269–278.