

MICROSCOPY AND MICROANALYSIS



CAMBRIDGE
UNIVERSITY PRESS

Detecting Clusters in Atom Probe Data with Gaussian Mixture Models

Journal:	<i>Microscopy and Microanalysis</i>
Manuscript ID	MAM-APTM16-16-146.R3
Manuscript Type:	APT&M2016 Special Issue
Date Submitted by the Author:	05-Feb-2017
Complete List of Authors:	Zelenty, Jennifer; University of Oxford, Department of Materials Dahl, Andrew; Department of Statistics Hyde, Jonathan; National Nuclear Laboratory Smith, George; University of Oxford, Department of Materials Moody, Michael; University of Oxford, Department of Materials
Keywords:	Atom probe tomography, Clustering, Gaussian mixture models, Expectation maximization

SCHOLARONE™
Manuscripts

IntroductionAbstract:

Accurately identifying and extracting clusters from atom probe tomography (APT) reconstructions is extremely challenging, yet critical to many applications. Currently, the most prevalent approach to detect clusters is the maximum separation method, a heuristic that relies heavily upon parameters manually chosen by the user. In this work, a new clustering algorithm, GEMA, was developed. GEMA utilizes a Gaussian mixture model to probabilistically distinguish clusters from random fluctuations in the matrix. This machine learning approach maximizes the data likelihood via expectation-maximization: given atomic positions, the algorithm learns the position, size, and width of each cluster. A key advantage of GEMA is that atoms are probabilistically assigned to clusters, thus reflecting scientifically meaningful uncertainty regarding atoms located near precipitate/matrix interfaces. GEMA outperforms the maximum separation method in cluster detection accuracy when applied to several realistically simulated datasets. Lastly, GEMA was successfully applied to real APT data.

1. Introduction

Atom probe tomography (APT), with its superior combination of chemical and sub-nanometer 3D spatial resolution, is an excellent tool with which to detect nanoscale precipitates (Gault et al., 2012). In many cases, particularly in the earliest stages of cluster formation, APT can identify features beneath the limits of most other conventional microscopy techniques. However, there exists a caveat: the user must independently extract the precipitates from the APT data ~~through the use of~~ using a cluster search algorithm. As the need to characterize clusters extends across many materials and users, a versatile and reliable cluster search algorithm is essential.

The analysis of nanoscale precipitates within data provided by APT is critical to many engineering applications in materials science. ~~As discussed previously, the~~ The nucleation and growth of Ni-Mn-Si precipitates can cause hardening and, ultimately, embrittlement, in ~~reactor pressure vessel (RPV)~~ reactor pressure vessel (RPV) steels (Wells et al., 2014, Styman et al., 2012). Additionally, the mechanical properties of Ni-based superalloys, which are used in jet engines and wind turbines, rely crucially on various solute additions and the resulting precipitation (Pollock et al., 2006). ~~Lastly~~ Furthermore, key research regarding oxide dispersion strengthened (ODS) steels focuses on the chemical characterization and distribution of oxide nanoparticles; although these highly stable particles are responsible for many of the exceptional mechanical properties of ODS steels, they are not yet fully characterized (London et al., 2015, Hirata et al., 2011). Hence, the need for accurate cluster identification and extraction of cluster information from APT reconstructions is imperative.

Numerous algorithms have been developed to identify individual clusters within APT data including the maximum separation method [7], (Hyde et al., 2000), the core-linkage algorithm (Stephenson et al., 2007), and Voronoi partitioning (Felfer et al., 2015). Currently, the most prevalent approach used to detect clusters is the maximum separation method. This heuristic relies heavily upon four parameters, each manually chosen by the user: Dmax, the maximum distance separating two solute atoms within the same cluster; Nmin, the minimum number of atoms a cluster must contain to be considered a cluster; and L and E, which are metrics related to the incorporation of solvent atoms within clusters. Although there have been

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

commendable efforts to provide guidelines for parameter selection and sensitivity analyses, the process still remains largely arbitrary and inconsistent (Hyde et al., 2011, Styman et al., 2013, Jäggle et al., 2014).

Formatted: Font color: Black

This raises the fundamental issue of scientific objectivity. Experiments should be reproducible, and results should be independent of the researcher. The application of current APT data [analysisclustering](#) algorithms requires selection of numerous user-defined parameters that unavoidably introduce subjective variation into the analysis. The outcome is a chronic lack of comparability between quantitative data obtained by different users. This represents a significant issue for APT-based research and becomes ever more apparent as the number of researchers in the field increases. Therefore, a more objective method for cluster analysis is essential.

Formatted: Font color: Black

In this work, a new cluster search algorithm, which uses a Gaussian mixture model (GMM) to probabilistically distinguish clusters from the matrix, is developed. This unsupervised machine learning algorithm maximizes the data likelihood via expectation-maximization (EM). Specifically, using APT data, the algorithm learns the position and size of each cluster. This new clustering algorithm is called the Gaussian mixture model Expectation Maximization Algorithm, or GEMA. Arguably, the most important feature of GEMA is that it eliminates the need for external parameter selection while simultaneously improving power, thus providing a key step towards routine reproducibility of quantitative measurements.

In this paper, k-means, one of the simplest machine learning algorithms used to cluster data, is first introduced to illustrate the key ideas underlying iterative clustering algorithms. Next the development of GEMA is described in detail, focusing on the interpretation of [GMMsGaussian mixture models](#) and how they are estimated using expectation maximization. The paper is concluded by illustrating the successful application of GEMA to both simulated and real APT data, respectively showing that GEMA is often superior to the state-of-the-art and that GEMA provides useful cluster representations of real APT data. Overall, this novel approach provides a powerful, statistically principled, and immensely extensible baseline for cluster detection in APT.

Formatted: Font color: Black

2. A first approach: k-means

Formatted: Font color: Black

k-means is one of simplest and most common clustering algorithms in statistics and machine learning (Bishop, 2006). After a random initialization, k-means jointly learns cluster locations and assigns atoms to clusters by iterating two steps: first, given cluster locations, each atom is assigned to the nearest cluster; second, given atom assignments to clusters, each cluster center is placed at the average location of all atoms assigned to that cluster (Figure 1). The value of this approach is that it breaks down a complex problem, i.e. jointly determining cluster parameters and assigning atoms, into two simple sub-problems, which are iteratively solved. GEMA uses a similar iterative approach, as do, more generally, all EM algorithms.

However, there are several disadvantages to using k-means on APT data. Primarily, all atoms must belong to a cluster in k-means, and thus atom assignments to the matrix are not possible. (Figure 2). Additionally, atoms must deterministically belong to a single cluster,

Formatted: Font color: Black

ignoring matrix/cluster and cluster/cluster uncertainty. Near precipitate interfaces, in particular, matrix/cluster uncertainty is a real scientific phenomenon which should be reflected in the modeling assumptions. Currently, all APT clustering algorithms share this undesirable property with k-means.

3. GEMA

Gaussian mixture models, (GMMs), which are essentially a more sophisticated alternative to k-means, can also be used to identify clusters. GEMA, which utilizes GMMs, overcomes the primary problem with k-means by explicitly modeling the matrix as an additional Gaussian component with infinite variance. A (bounded) Gaussian with infinite variance is identical to a uniform distribution, which formalizes the assumption that unclustered solute is distributed “randomly” in the matrix¹. GEMA addresses the latter shortcoming by probabilistically assigning atoms to clusters and the matrix, thus eliminating the strict cluster assignments imposed by k-means.

In regards to cluster assignments, GEMA labels atoms with a probability of 50% or greater as belonging to a cluster. Although the probability of an atom belonging to the cluster has been made visible to the user via atomic shading, this is not an arbitrary parameter. A threshold of 50% is used by default in the algorithm – as it is Bayes optimal in many contexts – and therefore cluster assignments remain consistent.

However, atomic probabilities can be extracted and utilized for additional analyses, such as cluster composition and cluster size (number of atoms). For instance, cluster composition could be determined by weighting each atom by its probability of belonging to the cluster.

Additionally, there are frequently potential gradients in chemistry and crystal structure around the precipitate boundary, owing simply to thermodynamics. Therefore, it is important to note that missing data, specifically at the boundary, could potentially affect cluster size. However, this problem is due to the nature of atom probe tomography as a technique, which lies outside the scope of this paper. In regards to chemical gradients, GEMA does not utilize the chemical identity of an atom when determining the probability of an atom belonging to a cluster. Therefore, chemical uncertainty does not affect cluster definition.

3.1. Gaussian Mixture Models

GEMA utilizes a Gaussian mixture model (Bishop, 2006) to probabilistically learn cluster locations and assign atoms to clusters and the matrix. As its name suggests, a Gaussian mixture model is a distribution defined by a weighted sum, or mixture, of Gaussians:

¹ In practice, this assumption is not necessarily true, nor does it need to be true in order for GEMA to be implemented effectively.

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

$$p(X) = \sum_{k=1}^K w_k N(X|\mu_k, \sigma_k^2) \quad (1)$$

where K is the number of Gaussian components and w_k , μ_k , and σ_k^2 are the weight, mean, and variance of the k -th Gaussian component, respectively. The goal of GEMA is to find a set of parameters, w_k , μ_k , and σ_k^2 for each k , such that this analytic distribution optimally matches the empirical distribution.

In its current state, GEMA assumes that the covariance matrix is spherical, or a multiple of the identity matrix. This is equivalent to the assumption that all clusters are spherical. However, despite this assumption GEMA works remarkably well on real APT data where this clearly isn't the case (see section 4.3). Furthermore, generalizing the covariance matrix to allow ellipsoidal clusters in a future version of GEMA would be relatively straightforward.

It should be noted that the current version of GEMA takes only solute atoms into consideration. Solute elements must be selected by the user prior to analysis. This solute element selection is the only input required by the user. All other parameters are either adaptively learned via the algorithm (μ_k , σ_k , w_k , and k) or hardcoded into the algorithm (τ).

3.2. Expectation-Maximization

An expectation-maximization (EM) algorithm (Bishop, 2006) was implemented to determine the most likely GMM for the given APT data. Similar to k-means, GEMA's EM algorithm alternates between an expectation (E) step and maximization (M) step until convergence.

First, the Gaussian parameters μ_k , σ_k , and w_k for each k are initialized (this is non-trivial and described in section 3.3). Next, each atom, n , is assigned a probability for being in each cluster, $p(k|x_n)$, based on its location, x_n . This is referred to as the E-step and can be described mathematically by the equation:

$$p(k|n) = \frac{w_k g(x_n; \mu_k, \sigma_k)}{\sum_{k=1}^K w_k g(x_n; \mu_k, \sigma_k)} \quad (2)$$

where $g(x_n; \mu_k, \sigma_k)$ is a D-dimensional Gaussian function. In this equation, the numerator is essentially a weighted probability of finding an atom, n , in cluster, k ; the denominator normalizes this probability across all clusters.

The Gaussian parameters are then optimized given the newly updated atomic probabilities of being in each cluster. This step, which maximizes the parameter likelihood, is called the M-step:

Formatted: Right

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Right

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

$$\begin{aligned}\mu_k &= \frac{\sum_{n=1}^N p(k|n)x_n}{\sum_{n=1}^N p(k|n)} & \sigma_k &= \sqrt{\frac{1}{D} \frac{\sum_{n=1}^N p(k|n) \|x_n - \mu_k\|^2}{\sum_{n=1}^N p(k|n)}} \\ w_k &= \frac{1}{N} \sum_{n=1}^N p(k|n)\end{aligned}\tag{3}$$

(4)

(5)

where N is the total number of atoms and D is the Gaussian dimension. Equations 3 and 4 determine the mean and standard deviation of cluster k, respectively; the contribution of each atom to these values are weighted by the probability of that atom being in the cluster, which was calculated in the previous E-step. Similar to Equation 2, the denominators of Equations 3 and 4 perform normalizations. Equation 5 computes the weight of a given cluster, k. This is accomplished by taking the probability of an atom belonging to cluster k and summing over all atoms.

The E-step and M-step are repeated until convergence², resulting in an estimate of μ_k , σ_k , and w_k for each cluster, k.

An illustration of an expectation-maximization algorithm being utilized to determine the best GMM to fit 2D data (with no background) is shown in Figure 3. In the first panel, (a), the data points are shown in green and the two randomly initialized Gaussians are shown in blue and red. The atoms are probabilistically assigned to the two clusters as shown in panel (b). Atoms are colored based on their probability of belonging to each cluster. Atoms that are equally likely to belong to either the blue or red cluster have been colored purple. This is referred to as the E-step. In panel (c), the Gaussian/cluster parameters (mean, weight, and variance) are updated (M-step) based on the cluster assignments from the previous step. In panels (d), (e), and (f) the E- and M-steps are repeated until convergence. A similar process is executed in GEMA, except for a few key differences: GEMA fits an additional background cluster; GEMA assumes the clusters are

² EM is guaranteed to converge to a local maximum of the likelihood.

spherical; and GEMA's initialization utilizes a kernel density estimate, which is described in the next section.

In summary, the solute clusters within the APT data are modelled using 3D Gaussian distributions, one for each cluster. GEMA learns the parameters of these Gaussians and determines which atoms belong to each one via an iterative process. In each iteration GEMA slightly improves its estimate as to which atoms belong to each cluster and the parameters of each Gaussian. These parameters include the location of the cluster (μ_k), the width of the cluster (σ_k), and the density of the cluster (w_k).

3.3. Initialization using Kernel Density Estimates

EM is an extremely useful tool; however, it is highly sensitive to initialization. Local maxima often prevent the algorithm from finding the global maximum, as convergence can occur at any local maximum. Fortunately, this problem can be solved with careful initialization.

Although initialization sensitivity is a thoroughly studied problem for general mixture models, typical initializations, such as k-means or repeated random restarts, were not viable for APT data. This is due to solute fluctuations within the matrix: initializations get stuck in these local maxima, which is not a problem for typical GMMs. Therefore, a novel initialization approach was developed specifically adapted to the peculiarities of APT data.

First, a kernel density estimate (KDE) (Bishop, 2006) was fit to the empirical solute distribution. KDEs are smoothed estimates of density; i.e., atoms from high solute density regions, notably from clusters, have high KDE values. The ~~location~~locations of the cluster centers were then initialized using this KDE. In particular, the solute atom with the highest KDE value was identified; this was the first initialization point. Using this point to initialize the cluster center, GEMA was run with K set equal to 1. GEMA's output was then used to produce a parametric estimate of atomic density, which was subsequently compared to the non-parametric density estimate given by the initial KDE. Successive initialization points were determined by subtracting the parametric density from the KDE and choosing the highest point, or, in other words, the point on the KDE which was not yet described by the model. This was done to ensure multiple initialization points did not come from the same cluster. This process was repeated until a specified number of initialization points were identified, and then stopped.

In other words, clusters were added one-by-one to the model to avoid fitting a single cluster with two Gaussians. As each cluster was modeled with a Gaussian distribution, the model was subtracted from the data, leaving the residuals from that cluster fit, as well as the remaining clusters. This new dataset was then fit with an additional Gaussian, and the process was repeated until all the clusters were identified.

The number of initialization points, or number of clusters, can be chosen via numerous model selection tools. One standard approach is to minimize the Bayesian information criterion (BIC) (Bishop, 2006). BIC is a model selection tool in statistics based on a compromise between the likelihood function and parsimony. The BIC is formally defined as:

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Indent: First line: 0.5"

Formatted: Font color: Black

Formatted: Space After: 12 pt

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

$$BIC = -2 \ln(\hat{L}) + k \ln(N), \tag{6}$$

where \hat{L} is the maximum of the likelihood function, k is the number of parameters being estimated, and N is the number of data points, in this case solute atoms. GEMA uses the BIC to automatically and objectively determine the optimal number of clusters (Figure 24).

4. Applications of GEMA

4.1. 2D Simulated Data

In this section, the capabilities of GEMA are visualized by considering several 2D simulated datasets. In Figure 35, GEMA is applied to a 2D APT simulation with two clusters, which are easily recovered. Atoms belonging to the matrix are drawn as hollow points, while the atoms belonging to clusters are color coded. Moreover, these atoms are shaded, with lighter shades indicating a lower probability of belonging to the cluster.

Figure 46 illustrates GEMA’s ability to recover a true cluster model given a noisy 2D simulated dataset. Each panel in Figure 46 is a kernel density estimate with height illustrating solute density at each point in 2D space. The true model, unknown to GEMA, is characterized by a uniform background with added Gaussian modes (left panel). Atoms are drawn from this true model distribution to produce a simulated two-dimensional APT dataset (center). Finally, GEMA is applied to the generated APT data, almost perfectly recovering the true model (right).

4.2. 3D Simulated Data

4.2.1. GEMA vs. Maximum Separation Method

Several simulations were used in order to test the accuracy of GEMA and compare it to the maximum separation method. The clustering performance of GEMA and the maximum separation method were quantitatively compared with Receiver Operating Characteristic (ROC) curves. ROCs plot the power, or true positive rate (the fraction of cluster atoms deemed in a cluster), against the false positive rate (the fraction of matrix atoms deemed in a cluster) to illustrate the trade-off between sensitivity and specificity presented by a method. Methods with a higher power outperform those with a lower power for a given percentage of false positives. As the choice of false positive level is subjective, a method is only considered superior if it has higher sensitivity for all levels of specificity.

The model estimates the probability of each atom belonging to a cluster, which is then thresholded at some level τ : if an atom has less than τ probability of being in a cluster, it is deemed to be in the matrix. This is the standard approach for ROCs. The maximum separation method, however, deterministically assigns atoms to clusters or the matrix; in other words, the choice of D_{max} (and other parameters) implicitly defines τ .

For GEMA, the ROC curves are created by smoothly varying the threshold τ from 0 to 1; for the maximum separation method, D_{max} is varied, thus enabling an unbiased comparison

between GEMA and the maximum separation method without the author subjectively choosing Dmax. The remaining three parameters for the maximum separation method were chosen in a consistent, standard manner: Nmin was set equal to the point at which the random cluster size distribution dropped to zero; L and E were set equal to Dmax.

The first simulation, previously detailed and utilized in (Moody et al., 2014), was characterized by a face-centered cubic structure in which sixteen precipitates were distributed uniformly at random. To model imperfect detection efficiency, 63% of the atoms were stochastically removed. Additionally, Gaussian noise was added to all atomic positions offsetting them from the lattice to reflect the imperfect resolution of APT. In this simulation, when considering only solute atoms, the clusters were approximately 72% more dense than the matrix.

GEMA was applied to the simulated dataset and accurately identified each of the sixteen clusters (Figure 57). More impressively, GEMA obtained a power, or percent of true positives, of 99.99% given a false positive rate of approximately 6%.

Due to the relatively simplistic nature of the precipitates in this simulation (Moody et al., 2014), both GEMA and the maximum separation method performed near-perfectly. The power of each method was comparable, with GEMA outperforming the maximum separation method by 0.2% at a 6% false positive rate. However, the clusters in Figure 57 all have similar cluster weights and widths; both of these properties, which are generally not realistic, favor the maximum separation method, which implicitly anticipates homogeneity between clusters³.

In order to illustrate the full capabilities flexibility of GEMA, two additional simulations were created: by varying the first varied the weight (w) distribution of each cluster weights (w) or widths (σ). Each simulation was created by drawing 1e4 solute atoms from a GMM where $k = 11$ (including the background cluster). In the first simulation (Figure 6) and the second contains relatively small clusters with an average σ of 8, the cluster widths were set equal to 2 (approximately 3nm 2.5nm in diameter) and the cluster weights varied between 1.4% to 2.6%. The next simulation had denser, smaller, and uniform weight clusters, where the cluster widths were equal to 0.8 (approximately 1nm in diameter) and the cluster weights were set to 0.4% (Figure 79). For these more complex precipitate distributions GEMA outperformed the maximum separation method by 18% and 66%, respectively, at a fixed false positive rate of approximately 3%.

It should be noted that in the third simulation, the ROC plot for the maximum separation method jumps sporadically as Dmax is varied. This illustrates that slight changes in Dmax can have drastic effects on the power when using the maximum separation method. This undesirable property is a product of having very sensitive parameters.

Lastly, a forth simulation was created in which cluster size was varied. In this simulation half of the clusters ($w = 2\%$ and $\sigma = 2.6$) were only approximately 25% denser than the matrix:

³ Typically, the same set of parameters are used to identify all clusters within a given dataset.

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Indent: First line: 0"

Formatted: Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Custom Color(34,34,34), Pattern: Clear (White)

Formatted: Font color: Black

Formatted: Font color: Black

($w = 80\%$), GEMA’s success in identifying these clusters shows promise in its ability to identify clusters within high solute concentration alloys (Figure 8–10). However, more thorough testing of the limits and capabilities of GEMA is the subject of on-going research.

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Indent: First line: 0.5"

4.3. Atom Probe Data

Formatted: Font color: Black

Next GEMA was applied to a model reactor pressure vessel (RPV) steel, provided by Rolls Royce Plc. Needle-shaped specimens of the steel were created from matchsticks via a standard two-stage electropolishing process (Gault et al., 2012). The samples were then analyzed via APT on a LEAP 3000X HR. The sample was run at 50K with a pulse fraction of 20%. The data reconstruction was performed using IVAS 3.6.8.

Formatted: Indent: First line: 0.5"

Small atomic density fluctuations do not significantly affect GEMA’s performance, as these fluctuations are effectively modeled by the background cluster. However, large crystallographic features, such as poles, were avoided for this analysis.

Due to detector efficiency, it is impossible to quantitatively determine the success of GEMA when applied to real APT datasets. However, by visual inspection, GEMA appears to successfully identify each of the clusters present (Figure 9–11).

Formatted: Font color: Black

For this sample, cluster identification was performed by GEMA on the order of minutes with a standard laptop (3.1 GHz). GEMA performs linearly with the number of solute atoms and quadratically with k . Therefore, run time for a given dataset depends on the percentage of solute atoms within the dataset as well the optimal number of clusters as determined via the BIC.

Formatted: Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font color: Black, Pattern: Clear (White)

5. Discussion

Formatted: Font color: Black

Formatted: Font color: Black

Accurately identifying and extracting clusters within atom probe datasets is critical to the field of APT. In order for the field to continue to grow as a reliable technique within the greater scientific community, cluster analysis results cannot vary from user to user; an algorithm that produces reliable and reproducible results is essential. GEMA can provide this reliability and reproducibility.

The three key virtues of GEMA are its principled probabilistic framework, generality, and extensibility. The probabilistic nature of this model has scientific merit, in that it quantifies genuine physical uncertainty regarding atoms at precipitate/matrix interfaces. By contrast, the maximum separation method simply provides an “in” or “out” verdict.

The generality of GEMA enables the algorithm to be applied to a wide range of APT datasets without requiring user input. This is illustrated in the first simulation where the maximum separation method performs very well for simple, homogeneous clusters, while GEMA excels in all studied circumstances.

Additionally, the modular nature of this approach facilitates modifications to the model. Potential extensions of GEMA are numerated in the following section.

5.1. Potential GEMA Extensions

GEMA currently models clusters as spherical Gaussians. Although this assumption does not noticeably hinder GEMA's performance, the covariance matrix could be generalized to allow ellipsoidal clusters in subsequent versions of GEMA.

Additionally, GEMA, in its current state, only considers solute atoms. If GEMA is to be used widely within the atom probe community, it would be necessary to incorporate a function which addresses non-solute atoms. The simplest extension would presumably automate the L and E step of the maximum separation method. Although seemingly straightforward, an intelligent automation is non-trivial.

In its current form, GEMA is limited to analyzing precipitates. However, this technique could presumably be adapted to analyze interfaces, such as grain boundaries, as well as other microstructural features. The probabilistic backbone of GEMA generalizes more flexibly than other state-of-the-art cluster detection approaches. Nonetheless, there exist several immediately apparent challenges. For instance, unlike precipitates, a grain boundary would likely be poorly modeled by a Gaussian distribution. This potentially could be circumvented by allowing the grain boundary to be modelled by multiple Gaussians and then joined post-hoc, or by including uniform slabs as mixture elements. Doing this correctly is likely possible and fruitful, but far from straightforward. Additionally, precipitates found along the grain boundary would likely require adding a deconvolution step to the algorithm. Consistently and correctly extracting grain boundaries and other microstructural features from APT reconstructions is an important challenge within the field, which seems possible to overcome via a machine learning approach similar to GEMA.

Another potential extension, and perhaps the most interesting, is that one could extend GEMA so that it learns not only the location, size, and width of each precipitate, but also information regarding phase. Through unsupervised learning the algorithm could simultaneously characterize different phases present in the dataset and determine the most probable phase of each precipitate. Furthermore, by adapting GEMA into a Bayesian approach the user could incorporate prior knowledge regarding phases commonly found within the material, as well as their level of certainty.

Finally, the algorithm could be extended to learn across APT datasets. One could, in principle, have a hierarchical model that learns the differences between materials at the top-level; differences between regions of the same material at the mid-level; and then, finally, run GEMA within each APT dataset while leveraging this information. Such hierarchical models are ubiquitous in the world of "big data" and extremely useful. As such GEMA provides a small but crucial first step toward integrating the trillions of data points generated from thousands of APT runs to make broad claims about the structure of many disparate materials.

6. Summary

In this [chapter paper](#), a new clustering algorithm, GEMA, was presented, which utilizes a

Formatted: Font color: Black

Formatted: Font: Not Bold, Font color: Black

Formatted: Font color: Black

Formatted: Indent: First line: 0.5"

Formatted: Font color: Black

Formatted: Indent: First line: 0.5"

Formatted: Font color: Black

Formatted: Font color: Black

Gaussian mixture model to probabilistically distinguish clusters from random fluctuations in the matrix. This machine learning approach maximizes the data likelihood via expectation-maximization: given atomic positions, the algorithm learns the position, size, and width of each cluster. A key advantage of GEMA is that atoms are probabilistically assigned to clusters, thus reflecting scientifically meaningful uncertainty regarding atoms located near precipitate/matrix interfaces. It was demonstrated that GEMA can outperform the maximum separation method in cluster detection accuracy when applied to realistically simulated data. Lastly, GEMA was successfully applied to real APT data.

7. Acknowledgements

[The EPSRC is kindly acknowledged for financial support under the grant EP/M022803/1.](#)

Formatted: Indent: First line: 0.5", No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font color: Black

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Cambridge, UK: Springer.

Formatted: Space After: 12 pt, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

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

Gault, B., Moody, M.P., Cairney, J.M. & Ringer, S.P. (2012). *Atom Probe Microscopy*. New York, USA: Springer.

Formatted: Font color: Black

Hirata, A., Fujita, T., Wen, Y.R., Shneibel, J.H., Liu, C.T. & Chen, M.W. (2011). Atomic structure of nanoclusters in oxide-dispersion-strengthened steels. *Nature Materials* **10**, 922-926.

Hyde, J.M. & English, C.A. (2000). An Analysis of the Structure of Irradiation induced Cu-enriched Clusters in Low and High Nickel Welds. Symposium R "Microstructural processes in irradiated materials," Fall MRS November, R6.6.

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font: Times New Roman, Font color: Black

Hyde, J.M., Marquis, E.A., Wilford, K.B., & Williams, T.J. (2011). A sensitivity analysis of the maximum separation method for the characterisation of solute clusters. *Ultramicroscopy* **111**, 440-447.

Jäggle, E.A., Choi, P., & Raabe, D. (2014). The Maximum Separation Cluster Analysis Algorithm for Atom-Probe Tomography: Parameter Determination and Accuracy. *Microscopy and Microanalysis* **20**, 1662-1671.

London, A.J., Santra, S., Amirthapandian, S., Panigrahi, B.K., Sarguna, R.M., Balaji, S., Vijay, R., Sundar, C.S., Lozano-Perez, S. & Grovenor, C.R.M. (2015). Effect of Ti and Cr on dispersion, structure and composition of oxide nano-particles in model ODS alloys. *Acta Materialia*. **97**, 223-233.

Formatted: Font color: Black

Moody, M.P., Ceguerra, A.V., Breen, A.J., Cui, X.Y., Gault, B., Stephenson, L.T., Marceau, R.K.W., Powles, R.C., & Ringer, S.P. (2014). Atomically resolved tomography to directly inform simulations for structure-property relationships. *Nature Communications* **5**, 1-10.

Formatted: Font color: Black

Pollock, T.M. & Tin, S. (2006). Nickel-Based Superalloys for Advanced Turbine Engines: Chemistry, Microstructure and Properties. *Journal of Propulsion and Power* **22**, 361-374.

Stephenson, L.T., Moody, M.P., Liddicoat, P.V., & Ringer, S.P. (2007). New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data. *Microscopy and Microanalysis* **13**, 448-463.

Styman, P.D., Hyde, J.M., Wilford, K., Morley, A. & Smith, G.D.W. (2012). Precipitation in Long Term Thermally Aged High Copper, High Nickel Model RPV Steel Welds. *Progress in Nuclear Energy* **86**, 86-92.

Formatted: Font color: Black

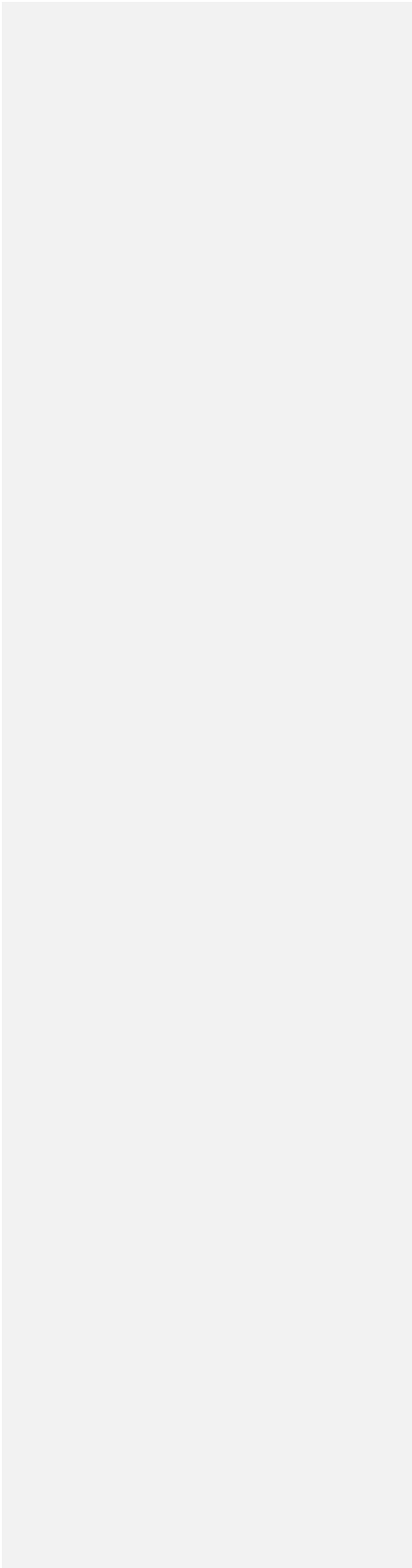
Styman, P.D., Hyde, J.M., Wilford, K., & Smith, G.D.W. (2013). Quantitative methods for the APT analysis of thermally aged RPV steels. *Ultramicroscopy* **132**, 258-264

Formatted: Font color: Black

Wells, P.B., Yamamoto, T., Miller, B., Milot, T., Cole, J., Wu, Y. & Odette, G.R. (2014). Evolution of manganese-nickel-silicon-dominated phases in highly irradiated reactor pressure vessel steels. *Acta Materialia* **80**, 205-219.

Formatted: Font: Times New Roman, Font color: Black

For Peer Review



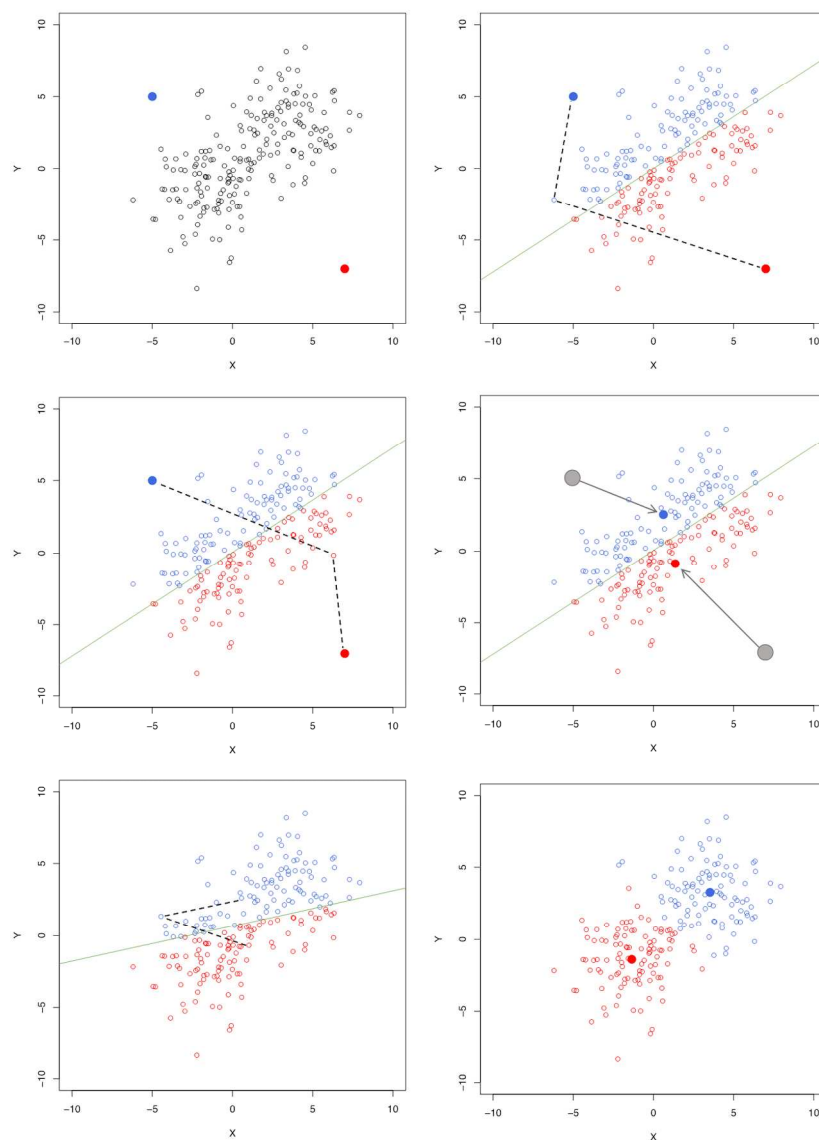


Figure 1: Illustration of how k-means assigns atoms to clusters. In (a) the two cluster mean locations are randomly initialized. (b) and (c) show how atoms are assigned to the nearest cluster mean (E-step). In (d), the cluster means are updated (M-step). The grey circles represent the original cluster means, while the arrows indicate the new, updated cluster means. In (e) and (f), respectively, the E-step and M-step are repeated until convergence.

353x506mm (300 x 300 DPI)

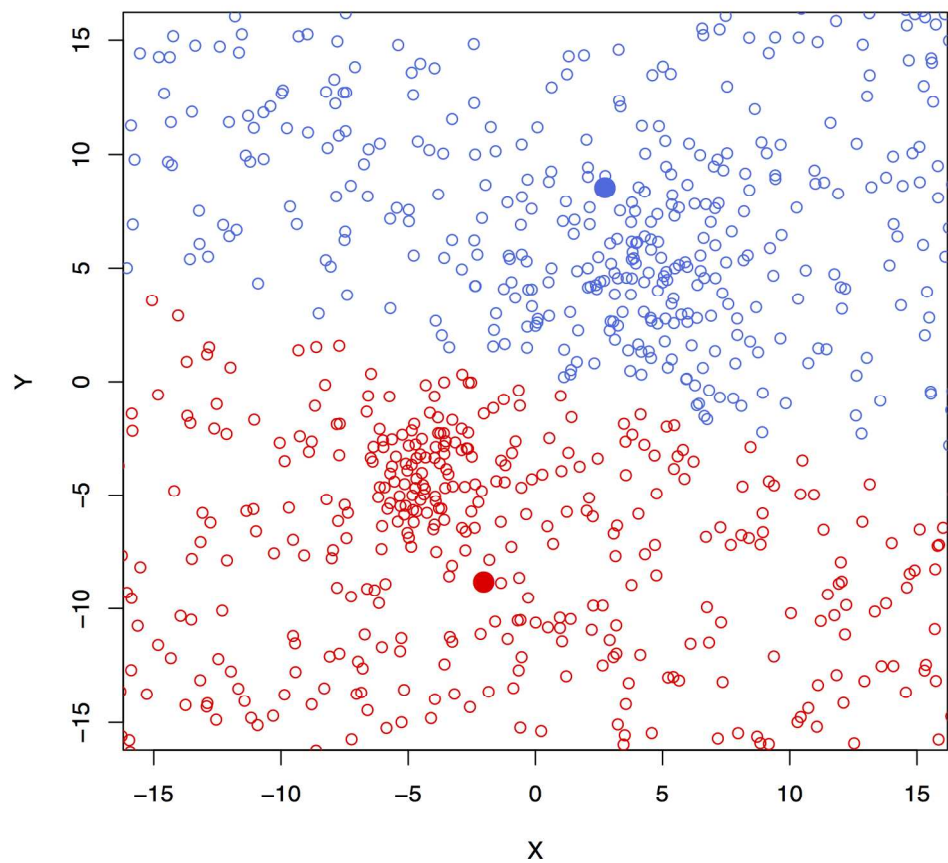


Figure 2: The output of k-means for a 2D dataset containing two clusters. Cluster assignments are shown in red and blue. The implicit requirement of k-means that all atoms must belong to a cluster lead to its failure in modeling solute atoms within the matrix.

177x177mm (300 x 300 DPI)

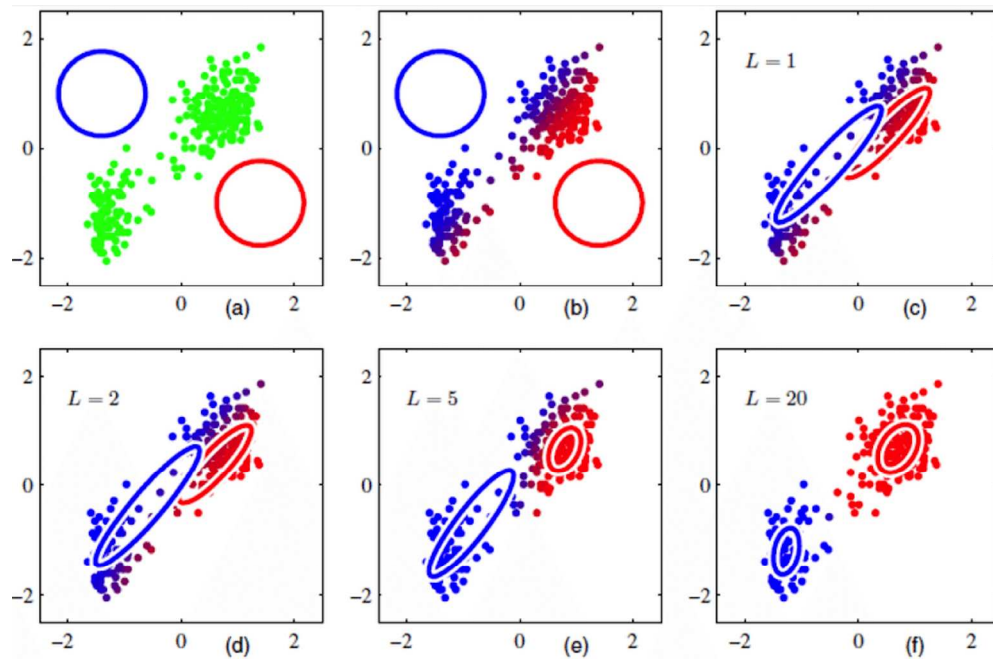


Figure 3: Illustration of an expectation-maximization algorithm being utilized to determine the best GMM to fit 2D data (Bishop, 2006). In (a) two Gaussian means are randomly initialized. The atoms are then assigned to the nearest cluster mean (b). This is referred to as the E-step. In panel (c), the cluster means are updated (M-step). In panels (d), (e), and (f) the E- and M-steps are repeated until convergence.

299x195mm (300 x 300 DPI)

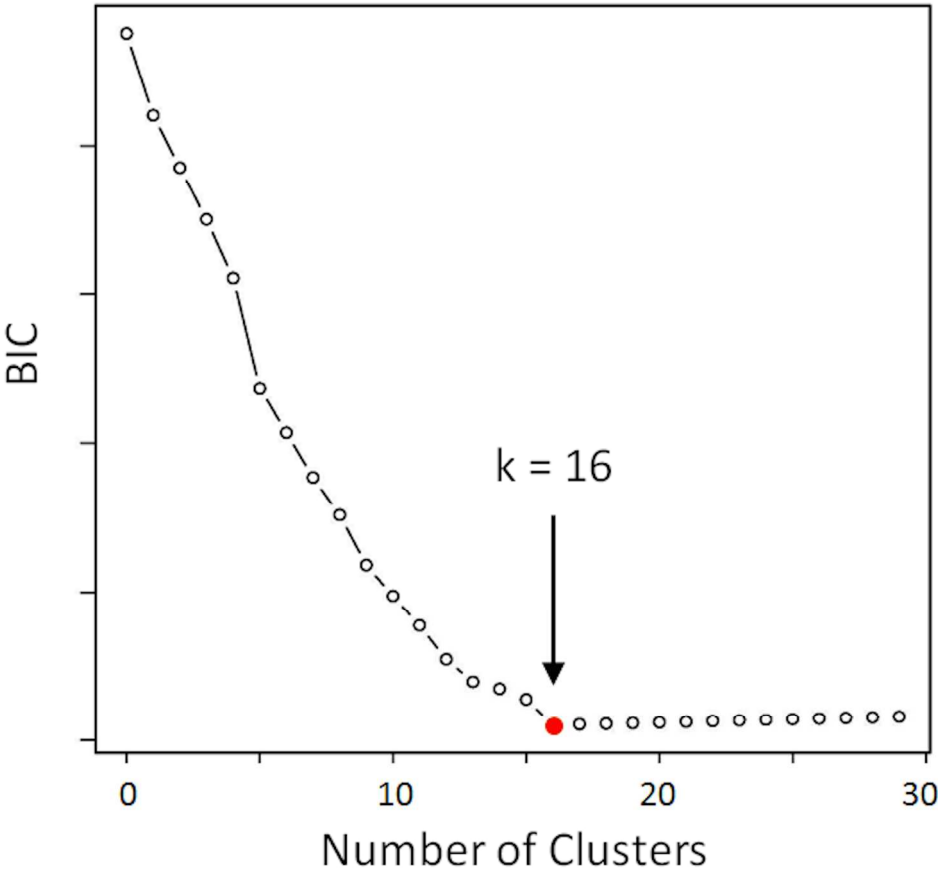


Figure 4: The number of clusters is determined by minimizing the BIC. The selected number of clusters (K) is highlighted red. Specifically, this BIC plot corresponds to the simulation in Figure 7.

171x164mm (300 x 300 DPI)



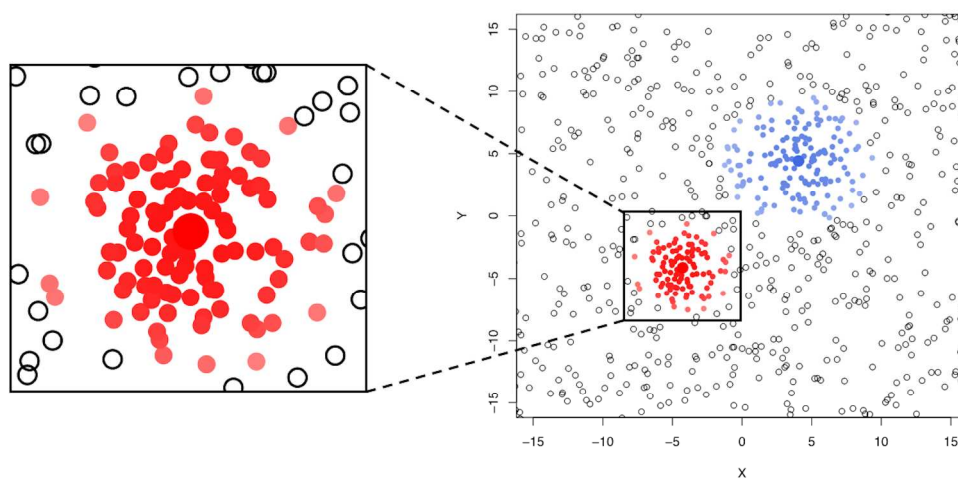


Figure 5: 2D atom probe simulation containing two solute clusters. Matrix and clustered atoms are indicated by hollow and solid dots, respectively. Shading indicates the probability that an atom is associated with a specific cluster.

324x164mm (300 x 300 DPI)

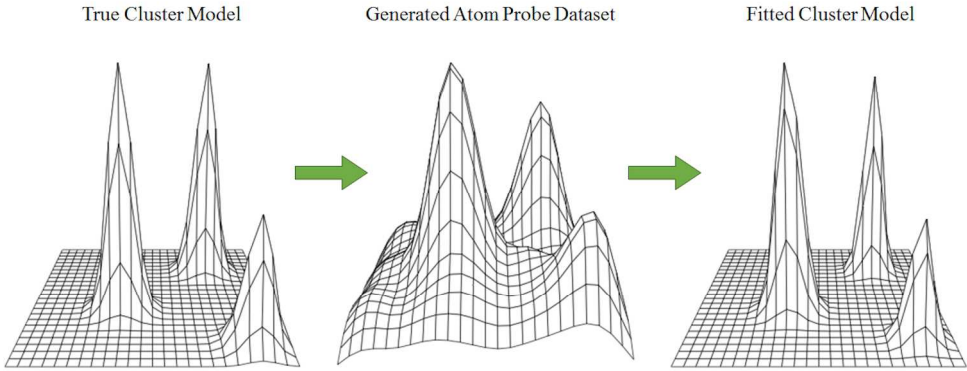


Figure 6: The true cluster model (left) was used to randomly generate a 2D atom probe dataset (center) containing three distinct clusters (these three clusters correspond to the three peaks seen in each panel). The fitted cluster model (right) shows the GEMA reconstruction. Each panel is shown as a kernel density estimate with height representing solute density.

352x142mm (300 x 300 DPI)

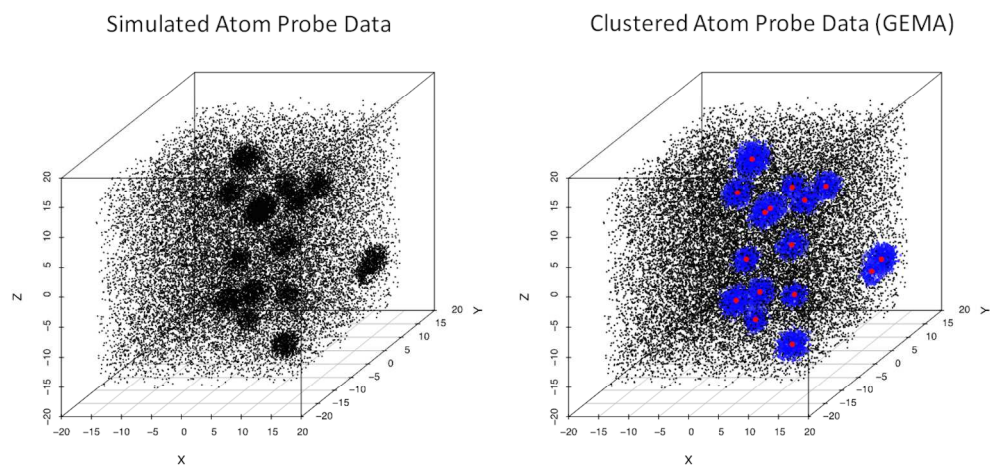


Figure 7: A simulated dataset (left) was used to test the accuracy of GEMA. The clustered dataset (right) highlights the clusters (blue) identified by GEMA. The learned cluster centers are shown in red.

334x161mm (300 x 300 DPI)

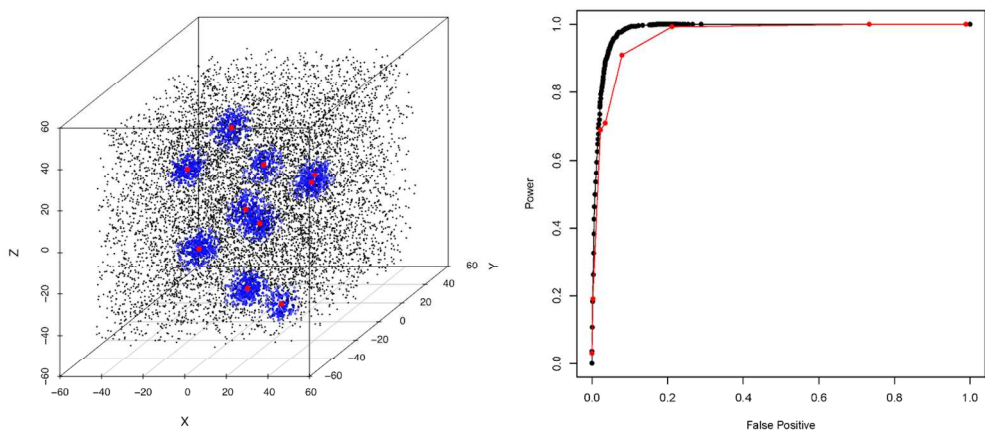


Figure 8: GEMA was run on the second simulation, which contained clusters varied in weight. GEMA probabilistically assigned atoms to clusters (blue) and the matrix (black). The learned cluster centers are shown in red. On the left, the ROC plot for this simulation is shown. GEMA is shown in black and the maximum separation method is shown in red.

362x166mm (300 x 300 DPI)

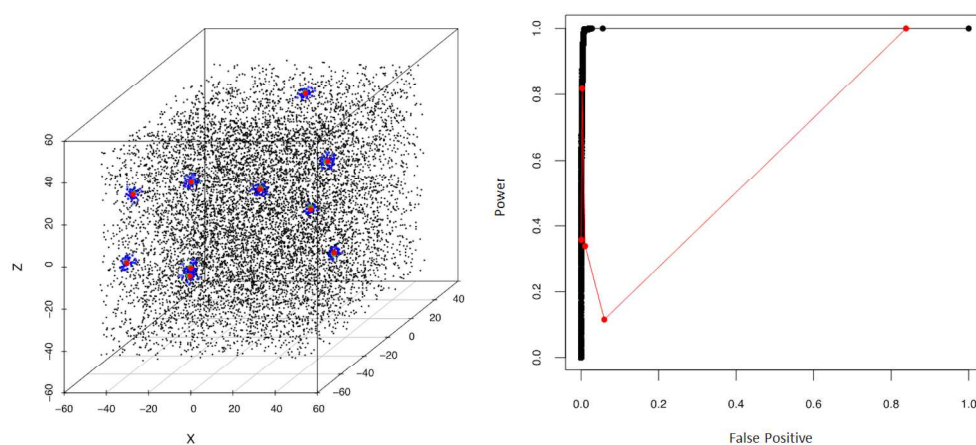


Figure 9: GEMA was run on the third simulation, which contained clusters with an average σ of approximately 3nm. GEMA probabilistically assigned atoms to clusters (blue) and the matrix (black). The learned cluster centers are shown in red. On the left, the ROC plot for this simulation is shown. Again, GEMA is shown in black and the maximum separation method is shown in red.

361x169mm (300 x 300 DPI)

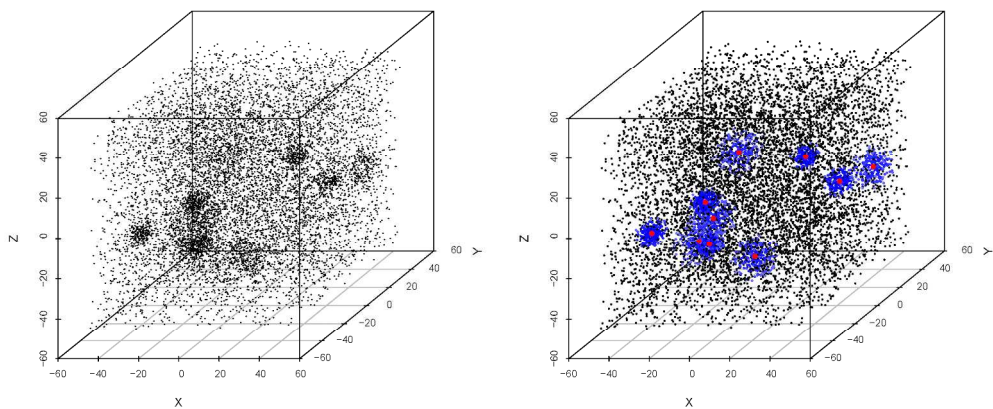


Figure 10: GEMA results for the fourth simulation in which cluster size was varied. GEMA probabilistically assigns atoms to clusters (blue) and the matrix (black). The learned cluster centers are shown in red.

726x317mm (300 x 300 DPI)

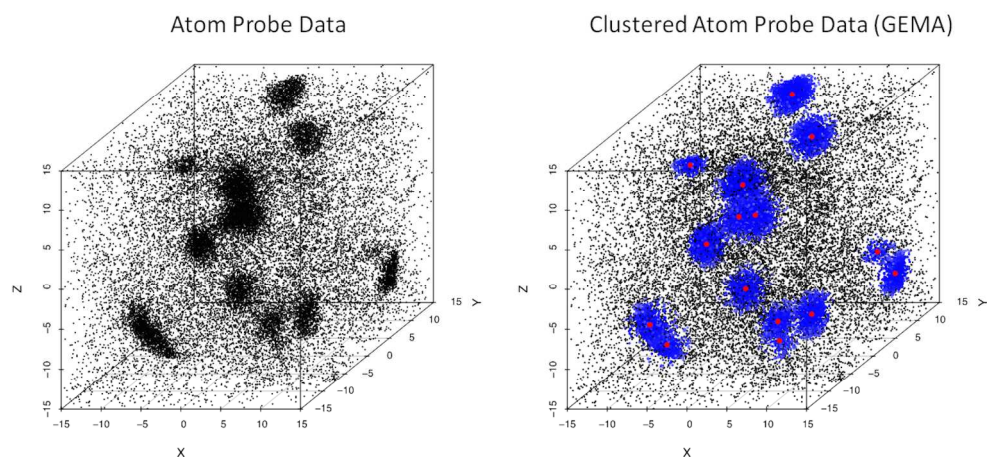


Figure 11: GEMA is run on raw atom probe data (left), giving the clustered output shown (right). GEMA probabilistically assigns atoms to clusters (blue) and the matrix (black). The learned cluster centers are shown in red. Although several of the clusters appear to be overlapping, they are, in fact, separate and distinct. This is simply an artifact due to the 2D projection of the 3D reconstruction.

335x158mm (300 x 300 DPI)