Visualising multi-dimensional structure/property relationships with machine learning

Baichuan Sun and Amanda S Barnard
Data61, CSIRO, Door 34 Goods Shed, Village St, Docklands VIC 3008, Australia
E-mail: amanda.barnard@data61.csiro.au

Keywords: machine learning, self-organising maps, silver, platinum, nanoparticles

Abstract
Data visualisation is an important part of understanding the distributions, trends, correlations and relationships in materials data sets, as well as communicating results to others. Traditionally visualisation has been straightforward, particularly when studying single-structure/single-property relationships. It is not so straightforward when confronted with a materials data set represented by a large number of features, and containing multi-structure/multi-property relationships. Here we use Kohonen networks, or self-organising maps, to aid in the visualise sets of silver and platinum nanoparticles based on structural similarity and overlay functional properties to reveal hidden patterns and structure/property relationships. We compare these maps to a popular alternative dimension reduction method and find them superior for our cases where the structure/property relationships are highly nonlinear, and the data set is imbalanced, as they often are in materials science.

Materials data is inherently multi-dimensional, with numerous quantitative features emanating from the various analytical and statistical techniques used today in computational and experimental materials science. It is relatively simple to compile an exhaustive list of descriptors during computational studies, and a comprehensive list during experimental studies that drawn on multiple characterisation tools. Materials data is, however, typically sparse and restricted to a limited number of observations, due to the high expense (in terms of time and resources) to capture each point. Materials researchers are generally confronted with a small data set (compared to fields such as image processing) with high dimensionality and many hidden nonlinear correlations [1–8].

Dealing with a small data set in multiple dimensions comes with unique challenges. The first is ensuring that the number of features does not exceed the number of data points [9], and that they are chosen and engineered appropriately [10, 11]. Feature engineering can assist in reducing the number of dimensions by eliminating features that are strongly correlated or by replacing groups of features with a new one, but this requires significant domain knowledge and may not be possible for all sets. In some materials data sets all of the features must be retained to provide a comprehensive description of the material. The next step is therefore to apply an appropriate machine learning algorithm to visualise high-dimensional data in a low-dimensional space, to facilitate a more straightforward comparison of structural features with properties. One such method is t-distributed stochastic neighbour embedding (t-SNE) which is a nonlinear dimension reduction technique developed by van der Maaten and Hinton [12] that models each high-dimensional object by a two- or three-dimensional point such that similar objects are adjacent and dissimilar objects are separated. This method is suitable for visualising high-level representations learned by an artificial neural network and has been applied to a variety of biomedical applications [13–15]. An alternative dimensional reduction method suitable for materials data set is a self-organisation map (SOM), or Kohonen network [16], which is an unsupervised artificial neural network [17] for nonlinearly mapping high-dimensional spaces into low-dimensional spaces that has the advantage of retaining the intrinsic topological relationship of the input data set. SOMs are also ideal for...
representing multi-dimensional information as a single two-dimensional plot to form a basis for further analysis [18], and have been recently used to create surface texture maps of nanoparticles for uses as materials fingerprints [19]. The applicability of these methods to multi-dimensional structure/property relationships of complicated sets of materials such as nanoparticles has yet to be established.

In this article we compare the use of these established methods (t-SNE and SOMs) to visualise multi-dimensional sets of metallic nanoparticles for the first time, to aid in the identification of underlying structure/property relationships hidden from traditional graphs used in materials and nanoscience. In each case we have used lists of materials features extracted from conventional statistical analysis and external labels based on domain knowledge, as is common practice in the field. As we will show, both methods are suitable to reduce the number of dimensions, but we find that SOMs have significant advantages that aid in interpretation and are more reliable for the types of data set common in materials science, avoiding possible misrepresentation.

Methods

The basic units of a SOM are neurons which are best organised on hexagonal grids. Hexagonal grids overcome anisotropy issues by including six neighbouring neurons, while other symmetries such as a square grids lack rotation invariance when counting more than four neighbours resulting in intrinsic anisotropy. All neurons can be arranged in a planar (approximate) rectangle in both cases, however the neurons on the edge will have smaller weights as they are far from all other cells in Euclidean distance, which reduces the symmetry of the resultant SOM. To overcome this issue, periodic boundary condition can be applied when calculating the Euclidean distance between neurons, to connect the upper and lower boundaries, and the right and left boundaries, so that the SOM plane occupies the surface of a torus. This is essentially the same as the conventional periodic boundary conditions that are used in simulations of bulk materials.

The weights of all neurons are initialised using random numbers. During each training step every neuron competes against all others until each original data point finds the one neuron that is closest to it in Euclidean space, referred to as the best matching unit (BMU). Given an input data point \( x \), and the weight of neuron \( i, j \) is \( w_{ij} \), then the Euclidean distance \( D \) is:

\[
D = \sqrt{\sum_{v=1}^{d} (x_v - w_{ij}^v)^2},
\]

where \( v \) is each component of the vector and \( d \) is the dimension of the standardised/normalised data set. Once the BMU is located, the weights of all neurons centred on it are updated. Initially all neurons on the SOM are updated, until only the BMU is updated at the conclusion of training. This ensures the radius of neighbourhood \( \delta \) is each component of the vector and \( d \) is the dimension of the standardised/normalised data set. Once

The total number of epochs, \( n_{\text{epoch}} \) is an important hyper-parameter of SOM training. A small number of epochs will result in under-training, and dissimilar data points may be adjacent on the final SOM reducing the resolution. A large number of epochs will results in over-training, leading to a waste of computational resources. It is possible to measure the number of void neurons which have no weight, and stop the training when this number stops decreasing over a threshold after, for instance, \( n_{\text{epoch}} = 5 \), to improve efficiency and autonomy [20].
As mentioned above, an alternative way to visualise high-dimensional data in a low-dimensional space is to use t-SNE. For \( n \) objects \( x_i \) in \( d \) dimension, the conditional probability that \( x_i \) would be picked as a neighbour given \( x_j \) is expressed as:

\[
P(x_i|x_j) = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_k \exp(-\|x_i - x_k\|^2 / 2\sigma^2)},
\]

(5)

where \( \sigma \) is the variance of the Gaussian that is centred on \( x_i \). \( \sigma \) is adapted to the density of the data such that smaller values are used for data in high density. The joint probabilities in the high-dimensional space is then set as:

\[
P(x_i, x_j) = \frac{P(x_i|x_j) + P(x_j|x_i)}{2n}.
\]

(6)

For the low-dimensional counterparts, \( y_i \) of the high-dimensional data \( x_i \), a Student t-distribution with a degree of freedom one is used to compute the joint probabilities \( Q(y_i, y_j) \) as

\[
Q(x_i, x_j) = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k (1 + \|y_i - y_k\|^2)^{-1}}.
\]

(7)

This distribution is heavy-tailed so that dissimilar objects in low-dimensional space can be modelled even if they are far apart. The optimisation objective is to have \( P(x_i, x_j) \) and \( Q(x_i, x_j) \) be equal, which means minimising the cost function \( \sum_{ij} P_{ij} \log(P_{ij}/Q_{ij}) \).

**Results**

Two computational data sets have been used to compare t-SNE and SOM materials dimension reduction capabilities. The first contains 425 silver nanoparticles with a diameter between 0.5 nm and 4.9 nm (13 to 2947 Ag atoms), and a range of different morphologies defined by zonohedrons enclosed by \{100\}, \{110\}, \{111\}, \{210\}, \{113\}, \{331\} and \{123\} surfaces. Readers should note that due to the the constraints imposed by the crystallographic lattice and the surface orientations characteristic of each zonohedron, it is not possible to generate each shape at all sizes, and the number of atoms in each nanoparticle is different. As described in [6, 19, 21] this set is characterised using a list of 24 features based on a logical set of descriptors common in computational and experimental materials science, including eight features capturing the local bonding configurations (examples of chemical information typically obtained computationally), six features capturing the bulk structures (examples of materials information that could be obtained from diffraction) and 10 features capturing surface structures (examples of nanoscale information that could be potentially obtained from spectroscopy). Any group of features could be chosen, but the aim here is to test our methods and so a diverse but intuitive list is desirable. The entire data set is available for download, including metadata listing the features and labels [22]. As shown in the supporting information (figure S1 available online at stacks.iop.org/JPMATER/2/034003/mmedia), this data set is both nonlinear in all the features, with imbalances in some features (bi-modal or tri-modal distributions) requiring stratification before classification and/or regression [6].

The results using both dimension reduction methods on this data set are illustrated in figure 1 coloured by an external label; the geometric shape of the nanoparticle. Here we can see that nanoparticles mapped together tend to be geometrical similar, but it is important to remember (particularly when interpreting the t-SNE plot in figure 1(a)) that these are not clustering algorithms, even though there are regions of the reduced space where similar structures (such as the Marks decahedra (18)) appear to be clustered together. It is also important to remember that the SOM (figure 1(b)) has periodic boundary conditions connecting the left and right sides, and the top with the bottom, while the t-SNE plot does not. In both cases it is easy to see patterns in the data via visual inspection, and to obtain some sense of how the shapes are related to each other, taking into account all of the structural features used to characterise the silver nanoparticles in the set. We have also encoded these maps (figures 1(c) and (d)) with the average nanoparticle diameter (in nm), which is the confounding variable (influencing both the dependent variable and independent variables) to which trends in nanotechnology are often attributed. Confounding plays an important role in statistical analysis seeking exact relationships, but is less important in machine learning which is concerned with identifying the ‘most likely’ value of a dependent variable given a set of predictors.

Based on the t-SNE results figure 2 shows a heat map used to encode the absolute values of charge transfer properties in electron volts. Here we can see that the ionisation potential (figure 2(a)), the electron affinity (figure 2(b)) and the electronic band gap (figure 2(c)) are consistently distributed over the space with variations in the centre across both structurally dissimilar groups. In the case of the energy of the Fermi level (figure 2(d)) we can see by comparing to figure 2(a) that there is a strong relationship to the external shape label. This
confirms that there is a shape-dependent structure/property relationship, even though we have not explicitly included the shapes as a descriptor [6], and is supported by more sophisticated machine learning methods published elsewhere [19]. In general twinned shapes have lower Fermi energies, whereas shapes with high \{111\} surface area have higher Fermi energies.

The same patterns are present in the SOMs when we encode the map with the electron transfer properties (figure 3). The electron affinity is low at the corners of the map (connected via periodic boundary conditions) where the ionisation potential and the band gap are highest. We know from comparison with figure 1(b) that this region is populated with shapes enclosed with a high fraction of \{111\} facets. The energy of the Fermi level is again strongly shape-dependent, with outliers clearly visible at the corners of the map, indicating that the outliers are also shapes with a high fraction of \{111\} facets. This is consistent with the high importance of the features (SCN6 and SCN9) characteristic of \{111\} facets to the value of the Fermi level identified in [6].

The second computational data set consists of 690 disordered platinum nanoparticles ranging in size from 1.5 to 7.5 nm in average diameter (165–15 837 Pt atoms). This set of disordered platinum nanoparticles was generated using combination of traditional molecular dynamics simulations and simple statistical data processing, to ensure the samples are physically realistic, thermodynamically stable and characteristic of nanoparticles observed experimentally, but guaranteed to be structurally unique. This data generation and cleaning process is described in detail elsewhere [23–25]. In addition to a list of 25 structural features this set also contains processing features extracted from the original molecular dynamics simulations, including the growth time (t), the growth rate (τ) and the growth temperature T. To capture the effects of inhomogeneous reaction kinetics and thermal fluctuations during formation and growth 70 combinations of τ and T were used within a range of values characteristic of experiments [26] and previous computational studies [27]. This data set is also

Figure 1. Two-dimensional (t-SNE, SOM) plots of the silver nanoparticle data set based on structural features and coloured by (a), (b) the external label of the nanoparticle shape, including the (1) cuboctahedron, (2) great rhombicuboctahedron, (3) hexakis octahedron, (4) icosahedron, (5) truncated cube, (6) rhombic dodecahedron, (7) modified truncated octahedron, (8) octahedron, (9) truncated rhombic dodecahedron, (10) rhombic-truncated octahedron, (11) small rhombicuboctahedron, (12) tetrahedron, (13) tetrakis hexahedron, (14) truncated octahedron, (15) triaxis octahedron, (16) trapezohedron, (17) Ino decahedron, (18) Marks decahedron and (19) simple decahedron; and (c), (d) the absolute value of the average particle diameter. A full list of the features is provided in the metadata accompanying the freely available data set [22].
available for download, including metadata listing the features [28], and as shown in the supporting information (figure S2), also exhibits nonlinearity and significant imbalances (bi-modal and tri-modal) distributions in most of the features [25].

The platinum nanoparticles in this set are very different from the silver nanoparticles used above, as there are no external shape labels, since the shapes cannot be described as regular zonohedrons. The Pt nanoparticles in this set are characterised by significant bulk and surface disorder, which has previously been quantified using two separate classification methods. To determine whether an atom is part of the surface layer, a radial shell of equidistant points was placed around the atom at a radius near the bond distance cut-off (minimum between the first and second peaks in the radial distribution function). At each of these points, a test atom was inserted and checked to see if it overlapped any real atoms within the system. If no overlap occurred, the atom was classified as a surface atom; else it was classified as bulk. This is described in detail in the supporting information of [25].

General bulk ordering was originally classified using the $q_{66}$ bond order parameter scheme [29]. Briefly, for an atom $i$ with neighbours $n(i)$, the local orientational structure is characterised by:

$$
\hat{q}_{66}(i) = \frac{1}{n(i)} \sum_{j=1}^{n(i)} Y_{66}(\vec{r}_{ij}),
$$

where $Y_{66}(\vec{r}_{ij})$ are the spherical harmonics related to the orientation of vector $\vec{r}_{ij}$ between atom $i$ and its neighbour $j$. With the restriction to $l = 6$ used traditional to probe hard sphere packing, a vector $\hat{q}_{66}(i)$ was assigned to each atom with the element $m = -6, \ldots, 6$ given by:

Figure 2. t-SNE plot of silver nanoparticle data set using structured features, and coloured by (a) the ionisation potential, (b) the electron affinity, (c) the electronic band gap, and (d) the energy of the Fermi level. The values of these charge transfer properties are absolute, in eV. A full list of the features is provided in the metadata accompanying the freely available data set [22].
where the looping is over the first nearest neighbours defined by a cutoff distance corresponding to the minimum between the first and second peak in the radial distribution function; in this case 3.4 Å. A quantitative comparison of the similarity in the orientational bonding environments between two atoms was obtained from the dot product \( \mathbf{q}(i) \cdot \mathbf{q}(j) \). The similarity coordination \( n_i(i) \) of atom \( i \) was defined as the sum of all first nearest neighbours having a dot product value exceeding 0.7. Highly ordered regions tend to have values of \( n_i(i) > 10 \) and can be represented as close crystalline packing [29]. For the most highly order regions \( n_i(i) > 10 \) ring analysis of the first nearest neighbour bonding network was used to classify the local atomic environment into face-centred cubic, hexagonal close packed, icosahedral, decahedral other ordered structures. In cases where the atoms did not conform to one of these local atomic environments they were classified as disordered. Icosahedral, decahedral and other ordered atoms were not present in sufficient quantity in this data set to be used reliably as structural features, so the following results have focussed on face centred cubic, hexagonal closed packed and disordered structures. The populations of each \( q_6 \) parameters between 0 and 12 were used as features in this study, along with the bulk packing classifications. More information on the bulk characterisation methods can be obtained from [24] and its supporting information.

The t-SNE and SOM plots for the disordered Pt nanoparticle data set are shown in figure 4 coloured by the three dominant bulk structural classes: face centred cubic (figure 4(a)), hexagonal close packed (figure 4(b)) and disordered (figure 4(c)). Comparison of these results clearly indicate that both methods identify groups of nanoparticles with strong similarity; in particular the face centred cubic and the disordered nanoparticles which have almost identical maps with the colours reversed. The hexagonal close packed structures coexist with the face centred cubic structure, but they are mostly mapped adjacent to the disordered nanoparticles, indicating similarity with disorder rather than neat twins or stacking faults in an otherwise ordered structure. In the t-SNE plots there is a suggestion of three clusters in the data set, but the SOMs (which has periodic boundaries) confirms that the nanoparticles with a high fraction of face centred cubic atoms are adjacent on the map in a distinct region (connected via the boundaries).

The surface order and disorder was characterised using the surface curvature and classified separately based on the surface coordination and angles of the surface atomic layer. The surface curvature for each surface atom

\[
q_6(i) = \frac{\mathbf{q}_{6\text{nn}}(i)}{\left(\sum_{m=0}^{6} |\mathbf{q}_{6\text{nn}}(i)|\right)^{1/2}},
\]

Figure 3. SOM of silver nanoparticle data set using structured features, and coloured by (a) the ionisation potential, (b) the electron affinity, (c) the electronic band gap, and (d) the energy of the Fermi level. A full list of the features is provided in the metadata accompanying the freely available data set [22].
Figure 4. (t-SNE, SOM) plot of platinum nanoparticle data set using structured features, and coloured by (a), (b) the particle volume, (b), (c) the fraction of face centred cubic packed atoms, (e), (f) the fraction of hexagonal close packed atoms, and (g), (h) the fraction of disordered packed atoms. A full list of the features is provided in the metadata accompanying the freely available data set [28].
was calculated from the displacement vectors with its first nearest neighbours. Considering an atom $i$ of coordination 4, with its nearest neighbours in a near planar configuration, one can define four angles and the three atoms defining those angles also define four planes. The surface normal to each plane are obtained by the cross product of vectors $\vec{i}j$ and $\vec{i}k$, which are used to define the average surface normal vector along with the average angle between this average vector and the four surface normal vectors. This average angle defines the surface curvature angle; compared to a planar configuration would give an angle of zero. Based on the coordination of the atoms, their bond angles, and their non-planar curvature, the surface packing can then be defined as $\{100\}$ when the curvature is $< 15^\circ$, the atomic angle between $70^\circ < \Theta < 110^\circ$ and the coordination

Figure 5. (t-SNE, SOM) plot of platinum nanoparticle data set using structured features, and coloured by (a), (b) the concentration of Surface Defects (surface coordination number 1–3, which are indicators for CO oxidation reactions), (c), (d) the concentration of Surface Microstructure (surface coordination number 4–7, which are indicators for oxygen reduction reactions), and (e), (f) the concentration of Surface Facets (surface coordination number 8–11, which are indicators for hydrogen evolution reactions and hydrogen oxidation reactions). A full list of the features is provided in the metadata accompanying the freely available data set \cite{28}.
number is 4; as \{111\} when the curvature is $< 15^\circ$, the atomic angle is between $40^\circ < \Theta < 80^\circ$ and the coordination number is 6; and as \{110\} when the curvature is $> 15^\circ$, the atomic angle is between $40^\circ < \Theta < 80^\circ$ and the coordination number is 6. The populations of surface curvature in the ranges $0^\circ$ to $10^\circ$, $10^\circ$ to $20^\circ$, $20^\circ$ to $30^\circ$, $30^\circ$ to $40^\circ$ and $40^\circ$ to $50^\circ$ were used as features in this study, along with the populations of different surface coordinations numbers (1, ..., 12) and the surface packing classifications. More information on these surface characterisation methods can be obtained from [24] and its supporting information.

An advantage of including characterisation of the surfaces of these Pt nanoparticles is that it provides information on both the structure and a functional property. It has been shown the surfaces of metal nanoparticle catalysts can be classified by the concentrations of different types of surface features using the surface coordination numbered surface atoms [30], which can be used as indicators of catalytic efficiency.
These classes have been previously termed *Surface Defects* for surface coordination numbers 1, 2 or 3, which are indicators for CO oxidation reactions; *Surface Microstructures* for surface coordination numbers of 4, 5, 6 or 7, which are indicators for oxygen reduction reactions; and *Surface Facets* for planar configurations with surface coordination numbers of 8, 9, 10 or 11, which are indicators for hydrogen evolution reactions and hydrogen oxidation reactions. The counting of surface atoms (reaction sites) is a developing area of research that shows great promise for guiding metal nanocatalyst design [31, 32], and these assignments have been shown to be suitable for investigating active sites on nanoparticle surface in the past [33–37].

In figure 5 we have applied colour maps to represent the absolute concentration of each surface property indicator in sites/mmol. From here we can see that *Surface Defects* (figures 5(a), (b)) do not reside on similar nanoparticles, but are more likely to be present on-nanoparticles with a high fraction of bulk disorder. *Surface Microstructures* (steps, kinks and terraces, figures 5(c), (d)) reside on nanoparticles with a high degree of crystallinity, and are likely to be associated with-nanoparticles containing a high fraction of hexagonal close packed atoms within the bulk. *Surface Facets* (planar or near-planar surfaces, figures 5(e), (f)) are associated with highly disordered nanoparticles, and visual inspection of these maps suggest that the roughened surfaces on these nanoparticles may form a class of their own. This is consistent with a recent report using more sophisticated classification and regression [25].

In addition to inspecting structure/property relationships we can also encode the t-SNE plots and SOMs with the processing features used in the molecular dynamics simulations to grow the nanoparticles, including $t$, $\tau$ and $T$, to uncover patterns relating growth conditions to the degree of disorder or the catalytic property indicators. These results are shown in figure 6, and by comparing with figure 5 we can identify some possible hidden process/property relationships by visual inspection. Figures 6(a), (b) we can see that the nanoparticles grown for longer periods of time are more likely to be decorated with *Surface Microstructures*, and have ordered interiors. Many of these nanoparticles will be the largest of the data set, but due to secondary nucleation occurring frequently at high growth rates (figures 6(c), (d)) this is not always the case. We can see a region of low $\tau$ in the upper right quadrant that corresponds to large $t$, occupied by the large nanoparticles, which are predominantly face centred cubic (see figures 4(a), (b)). High temperatures (figures 6(e), (f)) produces either face centred cubic or disordered nanoparticles, depending on $\tau$, but do not relate to a certain type of property indicator. Process/property relationships are strongly associated with $t$ and $\tau$ in this data set, but not $T$. This is also consistent with the findings of [25].

### Discussion

Both the t-SNE and SOM methods can reduce the dimensions and visualise patterns in nonlinear, multidimensional nanomaterials data sets, provided the data set has been standardised properly prior to training. Both methods can identify the possible existence of clusters, but should not be confused with clustering algorithms. Patterns appearing in t-SNE plots in particular can be misleading since t-SNE is a stochastic method and interpretation is not always straightforward. As mentioned above, t-SNE is a nonlinear algorithm that adapts to the underlying data, performing different transformations on different regions, which can be a major source of confusion. The size, shape, separation and density of the ‘clusters’ in figures 1(a), 2, 4(a), (c), (e), 5(a), (c), (e) and 6(a), (c), (e) are meaningless, making the SOMs in figures 1(b), 3, 4(b), (d), (f), 5(b), (d), (f) and 6(b), (d), (f) less susceptible to misinterpretation. The identification of clusters should always be undertaken with a proper clustering algorithm [38].

SOMs are unsupervised so they also do not require the user to apply categories (potentially based on incomplete knowledge or biased preferences) at the outset, which is well suited to complicated nanomaterials that cannot be easily labelled. In our cases the nanoparticles in the silver set could be assigned a shape but the nanoparticles in the platinum set could not. Machine learning methods amplify biases (particularly reinforcement learning methods), including researcher biases such as selection bias and confidence bias, which impact the choice of labels and hyper-parameters. SOMs have only the grid size, shape and the number of epochs to be predetermined, making it more robust against user biases. In contrast, t-SNE has many hyper-parameters to tune (such as the ‘perplexity’ that balances attention between local and global relationships, or others related to optimisation), which are often subjective and depend highly on the domain knowledge and technical experience of the user.

Another advantage of SOMs is that they can generate high resolution images, and are well suited to very large data sets. It is difficult to control the final resolution of maps generate using t-SNE, making them less suitable for creating ‘fingerprints’ needed for image processing. A disadvantage of SOMs, however, is that they are more computationally expensive.

At this point it should be pointed out that although both methods are designed to reduce high dimensional data to a lower dimension, both fail on very high dimensional data. In the case of materials data sets, since many
of the features are typically correlated [39], it is usually possible to reduce the number of dimensions prior to using t-SNE or SOM via feature engineering. If the number of dimensions is very large however, it is our experience that SOMs are more reliable. Without any prior knowledge, we recommend testing both methods in a series of computational experiments to inform the final method selection. More research will be required to universally validate these preliminary findings.

The next step having identified the patterns in the data using these methods is to test these relationships using classification and/or prediction algorithms. This has previously been reported for both data sets, using artificial neural networks to relate the features of silver with the energy of the Fermi level [6], and using random forest classifiers and regressors with extra trees to relate the structural features and the processing conditions to the property indicators of platinum [25].

Conclusions

We have presented here a comparison of two well established dimension reduction methods (t-distributed stochastic neighbour embedding and self-organising maps) applied to two different computational nanoparticle data sets, including a set of perfect silver nanoparticles characterised by their charge transfer properties, and a set of disordered platinum nanoparticles characterised by three classes of surface features known to be indicators of catalytic performance. We have shown that both methods can successfully reduce the multi-dimensional list of materials features and visualise the data in a convenient two-dimensional plot based on the similarity of the nanoparticles in the set. In each case researchers can clearly identify relationships between types of structures and properties via visual inspection, and in the case of Pt we extend this to relationships between the input ‘processing’ parameters such as growth rate, temperature and growth time.

Visual inspection of the SOMs presented in this work reveals a strong structure/property relationship between the shape of silver nanoparticles and the energy of their Fermi level, and a weaker relationship between shapes with a high fraction of \{111\} surface area and the ionisation potential, electron affinity and electronic band gap based on a large number of different types of features; even when the shape is excluded as an external label. Visual inspection of the SOMs presented for the Pt nanoparticle set also reveal some strong structure/property relationships as well as process/property relationships, such as the connection between the formation of Surface Microstructures with surface coordination numbers of 4, 5, 6 or 7 which are indicators for oxygen reduction reactions, and longer duration of synthesis (regardless of the growth rate), also associated with higher fractions of face centred cubic or hexagonal close packed atomic configuration. Shorter synthesis times mapped to regions associated with highly disordered nanoparticles, which are more likely to have highly coordinated surfaces with coordination numbers of 8, 9, 10 or 11, which are indicators for hydrogen evolution reactions and hydrogen oxidation reactions. The same data sets [22, 28] could be encoded by other labels or features to rapidly identify other patterns, and the same methods can be used to uncover possible the structure/property relationships worthy of more detailed analysis.

Both t-SNE and SOMs reveal underlying relationships in the data sets, which are most reliably interpreted with SOMs. In this study SOMs are preferred as their continuous distribution does not imply the presence of clusters (which may not exist), and is unaffected by imbalances in the data set such as an over-representation of one type of feature with respect to another (that can cause groupings easily misinterpreted as clusters). As they are unsupervised they are also less susceptible to researchers biases, and have fewer and less complicated hyper-parameters that need to be tuned. Both methods are suitable for use with experimental results as well as computational data, and may be useful in capturing patterns relating to more complicated synthesis, processing or operational conditions.

Acknowledgments

Computational resources for this project have been supplied by the Australian National Computing Infrastructure national facility under Grant q27, and Western Australia Pawsey Supercomputing Centre Athena project pawsey0238. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

ORCID iDs

Amanda S Barnard  https://orcid.org/0000-0002-4784-2382
References

[1] Agrawal A and Choudhary A 2016 *AIP Mater.* 4 053208
[2] Sun B, Fernandez M and Barnard A S 2016 *Nanoscale Horizons* 1 89
[3] Jain A, Hautier G, Ong S P and Perssson K 2016 *J. Mater. Res.* 31 977
[4] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 *Comput. Mater.* 3 54
[5] Ramakrishnan R and von Lilienfeld A 2017 *Rev. Comput. Chem.* 30 225
[6] Sun B, Fernandez M and Barnard A S 2017 *J. Chem. Inf. Mod.* 57 2413
[7] Ward L and Wolverton C 2017 *Curr. Opin. Solid State Mater. Sci.* 21 167
[8] Swann E, Sun B, Cleland D M and Barnard A S 2018 *Mol. Simul.* 44 905
[9] Ghiringhelli L M, Vybiral J, Levchenko S V, Dred C and Scheffler M 2015 *Phys. Rev. Lett.* 114 105503
[10] Ward L et al 2018 *Comput. Mater. Sci.* 152 60
[11] Wagner N and Rondinelli J M 2016 *Front. Mater.* 3 28
[12] Van Der Maaten L J P and Hinton G E 2008 *J. Mach. Learn. Res.* 9 2579
[13] Wallach J and Lilien R 2009 *Bioinformatics* 25 615
[14] Jamieson A R, Giger M L, Drukker K, Lui H, Yuan Y and Bhooshan N 2010 *Med. Phys.* 37 339
[15] Platzter A 2013 *PLoS One* 8 e56883
[16] Kohonen T 1998 *Neurocomputing* 21 1
[17] Bishop C 1995 *Neural Networks for Pattern Recognition* (USA: Oxford University Press)
[18] Gasteiger J, Li X, Rudolph C, Sadowski J and Zupan J 1994 *J. Am. Chem. Soc.* 116 4608
[19] Sun B and Barnard A S 2018 *J. Phys.: Mater.* 1 016001
[20] Wittek P, Gao S C, Lim I S and Zhao L 2017 *J. Stat. Software* 78 1
[21] Barnard A S and Sun B 2017 *Nanoscale* 9 12698
[22] Barnard A S and Sun B 2017 Silver Nanoparticle Data Set, CSIRO Data Collection (https://doi.org/10.4225/08/595f2a960c870)
[23] Barron H, Opletal G, Tilley R D and Barnard A S 2016 *Catal. Sci. Technol.* 6 144
[24] Sun B, Barron H, Wells B, Opletal G and Barnard A S 2018 *Nanoscale* 10 20393
[25] Sun B, Barron H, Opletal G and Barnard A S 2018 *J. Phys. Chem.* C 122 28085
[26] Ren T and Tilley R D 2007 *J. Am. Chem. Soc.* 129 3287
[27] Calvo F 2013 *Nanomaterials: From Fundamentals to Emergent Applications* (Amsterdam: Elsevier)
[28] Barnard A S, Sun B and Opletal G 2018 Disordered Platinum Nanoparticle Data Set, CSIRO Data Collection (https://doi.org/10.25919/9ba1831e866-4)
[29] Schilling T, Schöpe H J, Oettel M, Opletal G and Snoek I 2010 *Phys. Rev. Lett.* 105 025701
[30] Barnon H, Opletal G, Tilley R D and Barnard A S 2017 *Nanoscale* 9 1502
[31] Arons J, Jones L, Varambhia A, MacArthur K E, Ozkaya D, Sarwar M, Skylaris C K and Nellist P D 2017 *Nano Lett.* 17 4003
[32] Timperman L, Luo Y and AlonsoVante N 2016 *ChemElectroChem* 3 1705
[33] Falicov L M and Somorjai G A 1985 *Proc. Natl Acad. Sci. USA* 82 2207
[34] Spedendorf J S, Qinjin X, Goodpaster J D, Kenis P J A and Wieckowski A 2007 *J. Electrochem. Soc.* 154 F238
[35] Tian N, Zhou Z Y and Sun S G 2008 *J. Phys. Chem.* C 112 19801
[36] Zhao Z, Chen Z, Zhang X and Lu G 2016 *J. Phys. Chem.* C 120 28125
[37] van Santen R A 2017 *Modern Heterogeneous Catalysis: An Introduction* (Weinheim: Wiley-VCH)
[38] Xu D and Tian Y 2015 *Ann. Data Sci.* 2 165
[39] Feigl C, Motevalli B, Parker A J, Sun B and Barnard A S 2019 *Nanoscale Horiz.* accepted (https://doi.org/10.1039/C9NH00060G)