Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Catal. Sci. Technol.*, 2024, **14**, 6651

Unsupervised pattern recognition on the surface of simulated metal nanoparticles for catalytic applications†

Jonathan Y. C. Ting, *[a] George Opletal [b] and Amanda S. Barnard *[a]

Received 15th August 2024,
Accepted 21st September 2024

DOI: 10.1039/d4cy01000k

rsc.li/catalysis

The application of supervised machine learning to the study of catalytic metal nanoparticles has been shown to deliver excellent performance for a range of predictive tasks. However, this success assumes that the particles have been thoroughly characterised and that the property labels are known. Even in exclusively computational studies, the labelling of metal nanoparticles remains the bottleneck for most machine learning studies due to either high computational costs or low relevance to the experimental properties of interest. To facilitate more widespread use of machine learning in catalysis, a computationally affordable strategy to describe metal nanoparticles by a label that is relevant to their catalytic activities is needed. In this study we propose an entirely data-driven approach that can be automated to characterise the patterns and catalytic activities of the surface atoms of simulated metal nanoparticles, and evaluate its utility for catalytic applications.

1 Introduction

Heterogeneous catalysis is a cornerstone of industrial chemical reactions due to its high efficiency and selectivity.[1] The contribution of surface atoms toward the catalytic performance of heterogeneous catalysts is known to be multivariate.[2,3] The unique coordination environments, local symmetry, and electronic structures of these atoms directly influence the adsorption and activation of reactant molecules, and hence the subsequent reaction pathways.[4–6] Consequently, the ability to recognise the patterns among the surface atoms that give rise to good catalytic performance is a fundamental prerequisite for researchers to understand and optimise the catalytic behaviour of metal nanoparticles.

However, past attempts to characterise the surface of metal nanoparticles, which is a type of versatile and highly active heterogeneous catalyst, often approximate the catalytic contribution of surface atoms based on a single physically-meaningful variable. A notable example is the characterisation approach for metal nanoparticles inspired by the biological genomes, proposed by the Baletto group.[7–10] The method sequences the structural genome of metal nanoparticles based on a chosen atomistic geometrical variable to distinguish, catalogue, and count the adsorption sites available on the surface.[7] While the approach is general and could be used with other variables, the limitation of such approaches is that only the information in a single variable can be utilised. More accurate electronic variable from quantum mechanical calculations could potentially be used in place of the geometrical variable, but it would be prohibitively expensive for metal nanoparticles. Given the multivariate contributions of surface atoms toward the net catalytic efficiency of metal nanoparticles, the accuracy of attempts to predict the catalytic contribution of each surface atom using a single geometric variable will always be limited. A more judicious approach to capture information relevant to catalysis is to combine the information from multiple computationally affordable and catalytically relevant variables.

We propose that by framing the problem of recognising surface patterns as an unsupervised clustering task, surface atoms can be grouped based on their similarity in the high dimension encoded by multiple variables that are relevant to catalysis. This allows more complex and nuanced surface patterns to be identified, which can then be collectively related to a chosen variable that is known to be predictive of catalytic properties to identify applications that the groups are suitable for.

Some atomistic variables that have been used to study the catalytic properties of metallic systems include the orbital-wise coordination number,[11] effective coordination,[12] generalised coordination number (GCN)[13] and its variants.[14] Considering computation simplicity and correlation with catalytic performance, a good variable to evaluate the

[a] *School of Computing, 145 Science Road, Australian National University, 2601 Canberra, ACT, Australia. E-mail: jonathan.ting@anu.edu.au, amanda.s.barnard@anu.edu.au*
[b] *Data61, Commonwealth Scientific and Industrial Research Organisation, 3008 Melbourne, VIC, Australia*

† Electronic supplementary information (ESI) available. See DOI: **https://doi.org/10.1039/d4cy01000k**

catalytic relevance of the groups of metal nanoparticle surface atoms would be GCN, which is both simple and predictive,[15] and was also chosen for the aforementioned genomic approach.[7] It is expressed as:

$$\mathrm{GCN}_i = \sum_{j=1}^{N_i} \frac{\mathrm{CN}_j}{\mathrm{CN}_{\max}} \quad (1)$$

where CN_j is the coordination number of atom j, N_i is the number of first shell neighbours of atom i, and $\mathrm{CN}_{\max}$ is the maximum value among the coordination numbers of the neighbours of atom i.

According to the Sabatier's principle, there is an optimal bond strength that best catalyses a given reaction,[16] and GCN was deemed to be a useful variable to identify this optimal value.[15] Since its proposal, GCN has been applied in studies involving different metal nanoparticles and chemical reactions, including oxygen reduction reaction (ORR) catalysed by platinum (Pt) and gold (Au) nanoparticles,[8,17–19] carbon dioxide reduction reaction catalysed by copper (Cu) nanoparticles,[9,20,21] acetone reduction reaction (RCORRR) catalysed by Pt nanoparticles,[22] carbon monoxide oxidation reaction (COOR) catalysed by Pt nanoparticles,[23–25] and reverse water-gas shift reaction catalysed by Cu nanoparticles.[26]

While clustering methods have been employed to study collections of entire nanoparticles in the past,[27] studies clustering the individual atoms within nanoparticles are rare. In the work of Zeni *et al.* which characterised the melting of Au nanoparticles,[28] six classes of local atomic environment types were defined from a small database of configurations randomly extracted from the phase change trajectories of their simulations, using a hierarchical *k*-means clustering approach. Each atom was described by 40-dimensional features generated using a modified version[29] of the 3-body local atomic cluster expansion descriptor.[30] While the results concluded that Au cuboctahedra start to melt from the surface when heated, no conclusion related to catalytic performance was drawn.

In this article, we demonstrate the feasibility of using a clustering method to identify groups of patterns on metal nanoparticle surfaces, which are then evaluated based on a physically meaningful variable such as GCN to produce catalytically-relevant labels. In the following section, we explain the methodology to extract atomistic features from the raw nanoparticle coordinates, to preprocess the data, to cluster the surface atoms, and to evaluate the clustering results. This entirely data-driven method can be applied to the surface atoms of any metal nanoparticle, and the results for palladium (Pd) nanoparticles presented here show that the methodology allows for reliable separation between bulk and surface atoms and for subsurface layers of ordered nanoparticles to be identified, in addition to recognising the patterns among the surface atoms. Combining the visualisation of the feature profiles of these groups of atoms with the proposed evaluation metrics also enables researchers to further understand the characteristics of surface atoms that contribute to the catalytic performance toward chemical reactions of interest.

2 Methods

2.1 Data set

A set of simulated Pd nanoparticles are used as a testbed in this study. Idealised (sampled from simulation temperature of 0 K) polyhedra with eight different shapes, including the cuboctahedron (CO), cube (CU), decahedron (DH), icosahedron (IC), octahedron (OT), rhombic dodecahedron (RD), tetrahedron (TH), and truncated octahedron (TO), are used to develop the computational pipeline. This pipeline is then applied to slightly heated and disordered (DIS) nanoparticles to test its performance on more realistic specimens. The testbed contains a total of 39 nanoparticle structures, as tabulated in Table 1.

The testbed was taken from a Pd nanoparticles data set that was not produced as part of this study, but is publicly available.[31] This data set consists of 4000 Pd nanoparticle conformations generated from classical molecular dynamics simulations with embedded atom interatomic potentials,[32] ranging in size from 137 to 16 262 atoms (1.4 to 7.5 nm in diameter). The data set is diverse, and each structure is unique, including ordered crystalline nanoparticles, polycrystalline, and twinned nanoparticles, along with disordered and non-crystalline nanoparticles, depending on the growth temperature, growth rate, and simulation duration.

The raw data are transformed into potentially useful structural features that are more amenable to clustering algorithms at the atomistic level using software packages including the Network Characterisation Package (NCPac),[33] Atomic Simulation Environment,[34] Python Structural Environment Calculator,[35] and SYMMOL.[36] The features are grouped into five groups of descriptors, namely positional, geometric, Steinhardt, neighbour, and order descriptors. Detailed explanations of the features and the software parameters used to generate them are provided in ESI† (section S1).

2.2 Data preprocessing

The extracted and engineered features for each data set are preprocessed in the following ways to reduce redundancy:

1. Features with variance of 0.0 are removed.

Table 1 Descriptions of the simulated nanoparticles used as a testbed

Shape	Temperature (K)	Sizes	Total
CO	0	1	1
CU	0	3	3
DH	0	3	3
IC	0	1	1
OT	0, 323, 523	3	9
RD	0, 323, 523	3	9
TH	0, 323, 523	3	9
TO	0	1	1
DIS	723	3	3

2. One of each pair of features with Pearson correlation coefficient above 0.9 is removed. The features are ranked according to the ease of interpretation and relevance to catalysis, and the feature with lower ranking scores in each highly correlated pairs is retained. The ranking scores are included in ESI† (section S1.5).

3. Features are scaled using min-max normalisation.

4. Principal component analysis is conducted on the data to reduce the dimensionality, as described in ESI† (section S2.1). The number of features is set to be the number of components that retains >99% of the data variance.

2.3 Iterative label spreading

The iterative label spreading (ILS)[37] clustering algorithm was chosen for this work as it offers a few advantages over other alternatives,[38–40] such as no requirement for preliminary estimation nor optimisation of hyperparameters. ILS has also demonstrated greater reliability in both simple and challenging clustering tasks, including the null and chain cases, and was shown to be ideal for noisy datasets with high dimensionality and high variance, which are typical in materials science.[41–43]

ILS sequentially orders data samples, based on their proximity to initialised samples in a high dimensional space.[37] The initialisation process is described in ESI† (section S3.1), and the criterion of the labelling process is described by:

$$R_{\min}(i) = \min(\{r(i,j) | i \in L \text{ and } j \in U\}) \quad (2)$$

where $r(i, j)$ is the similarity distance metric between a labelled sample i and an unlabelled sample j, L is the set of all labelled samples, and U is the set of all unlabelled samples.

ILS returns an ordered minimum distance ($R_{\min}(i)$) plot when all samples are labelled, where i indicates the order by which the atoms are labelled (based on their proximity to the atoms that are already labelled in the feature space), and $R_{\min}$ reports the distance between the previously labelled point and the newly labelled point. The range of the plot thus corresponds to the number of atoms being clustered. The plot captures useful information such as the number of clusters and their separation in Euclidean space based on a series of peaks which signify drops in density between clusters. The clusters can be estimated by dividing the $R_{\min}$ plot at each peak into separate regions, and definitively identified by relabelling a sample in each region and reapplying ILS to assign the final cluster labels. Our method for automatically identifying peaks is described in ESI† (section S3.2). Double confirmation can be obtained by applying ILS to each cluster to ensure there are no hidden sub-clusters that could be further divided.

2.4 Cluster evaluation

The clustering outcome is internally evaluated quantitatively using the Silhouette coefficient,[44] Calinski–Harabasz index,[45] and Davies–Bouldin index,[46] since ground truth labels are not available (particularly for disordered nanoparticles). At a high level, Silhouette coefficient measures the similarity between a sample to its own cluster (cohesion) compared to other clusters (separation). A mean coefficient value of −1 across all samples indicates poor clustering performance, 0 indicates overlapping clusters, and +1 signifies highly dense clustering and better defined clusters. The Calinski–Harabasz index or Variance Ratio Criterion computes the ratio of the between-clusters dispersion over the within-cluster dispersion, dispersion being sum of distances squared, meaning a higher value represents denser and better separated clusters. On the other hand, Davies–Bouldin index denotes the average "similarity" between each cluster and the cluster it is most similar to, with similarity being measured by the distance between clusters with the size of the clusters themselves. Therefore, an index value closer to 0 signifies a better partition from clustering. The mathematical details of these methods are provided in ESI† (section S4).

In addition to these internal evaluation metrics, some domain-relevant cluster evaluation metrics have also been proposed based on the activity maps that depict the onset potential (potential in an electrochemical cell that drives the reaction) as a function of GCN, obtained from catalysis studies of different monometallic nanoparticles for different chemical reactions.[17,20,22,23] The aforementioned optimal bond strength that best catalyses a given reaction according to the Sabatier's principle can be reasonably represented by the peak of the activity map, which is composed of multiple linear equations. For example, a GCN of ~8.3 was found to be optimal for ORR catalysed by Pt nanoparticles,[17] while a GCN of 3.1 has been found to be optimal for the reduction of carbon dioxide to methane accelerated by Cu nanoparticles.[20] Many studies have investigated the structural characteristics of the atoms near the optimal GCN values to conduct surface engineering such that the nanoparticles contain more of these atoms.[17,18,20,22,23] Building upon this, we propose that the GCN activity map can also be utilised to evaluate the catalytic contribution of a given group of atoms.

A catalytic weighting profile (q) as a function of the GCN value can be obtained from these activity maps by normalising the maps such that the areas under the lines sum to 1. The normalisation is necessary for the evaluation metrics proposed below to be compared with each other as the intercepts of the equations directly affect the outcomes. The GCN distributions of each cluster (p) can then be compared with this profile for the cluster to be evaluated meaningfully. Further details about the activity maps and their normalisation are included in ESI† (section S5).

The evaluation metrics proposed here are termed as selectivity (eqn (3)), specificity (eqn (4)), and sensitivity (eqn (5)), and are designed to be bound within [0, 1] for ease of interpretation. These metrics are illustrated in Fig. 1, and defined as:

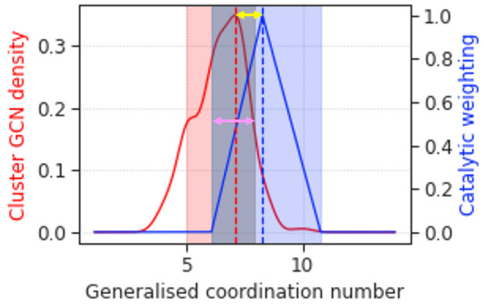

Fig. 1 Illustration of domain-relevant cluster evaluation metrics. The red and blue curves represent the kernel density estimation of the generalised coordination number distribution for a given surface cluster and the catalytic weighting profile obtained from the reference activity map for a chemical reaction of interest, respectively. Selectivity, specificity and sensitivity are defined by the yellow range, the pink range, and the area under the red curve within the pink range, respectively.

$$\mathrm{SEL} = 1 - \frac{\left|\arg\max_g p(g) - \arg\max_g q(g)\right|}{\max(G) - \min(G)} \tag{3}$$

$$\mathrm{SPC} = \begin{cases} -\dfrac{1}{2q(g)O(p,q)} + 1, & \text{if } p \text{ within } q \\ \dfrac{1}{2q(g)O(p,q)}, & \text{otherwise} \end{cases} \tag{4}$$

$$\mathrm{SEN} = \int \frac{p(g)q(g)}{\max(q(g)) - \min(q(g))}\,dg \tag{5}$$

where g is GCN, $p(g)$ is the GCN distribution of a given cluster, $q(g)$ is the catalytic weighting profile obtained from the reference reaction GCN-activity map, G is the range of GCN considered (1 to 14 for this work), and $O(p, q)$ is the overlapped GCN range between $p(g)$ and $q(g)$. Further details regarding the computation of $p(g)$, $q(g)$, and $O(p, g)$ are included in ESI† (section S6).

Selectivity denotes the difference in the GCN values corresponding to the peaks of both p and q. This informs how selective are the surface atoms toward the reaction of interest, which is important for the process of catalyst design. It is maximised when the peaks overlap, and minimised when each peak is at the extreme boundaries of GCN range considered. Specificity is related to the overlapping range of the distributions, and quantifies the exclusion of unwanted reactions. It is maximised when the full width at half maximum of p is entirely within q, and minimised when there is no overlap between them. We relate the metric termed sensitivity to the area of the overlapping distribution, which informs the proportion of the surface atoms that are actually useful for the catalysis of the reaction of interest,[47,48] which is deemed essential for the manufacturing process. It is maximised when the whole cluster comprises atoms with the reference reaction optimal GCN value, and minimised when there is no overlapping GCN range.

We note here some potential limitations of GCN: (i) It was warned that the analyses of GCN presuppose that there are no significant surface reconstructions upon adsorption, such that the geometric and electronic structures of the clean active sites are representative of those with adsorbates.[15] While this is a fair approximation in many cases, there can be exceptions for strong chemisorbates and/or large surface coverage of species.[49] (ii) GCN focuses on the geometric arrangement of atoms, neglecting electronic structure effects that might be crucial for understanding reactivity and catalytic performance. (iii) For multimetallic catalysts beyond bimetallic nanoparticles, GCN has not been proven to be able to sufficiently account for the interactions between different metal species and their collective impact on catalysis.[7] (iv) The calculation of GCN depends heavily on the identification of nearest neighbours, which is often based on a radial cutoff distance. While the cutoff distance is found to be robust even in highly disordered monometallic nanoparticles, the inclusion of other metal elements in multimetallic nanoparticles may cause overlapping of the first and second nearest neighbour peaks due to the mismatch between chemical species, and hinder accurate calculation of surface atom GCN.[7]

Nevertheless, similar to the work of Baletto group,[7] while GCN is used here for the evaluation of catalytic relevance of the surface clusters, the methodology is transferable to any other suitable variable that is capable of predicting the catalytic performance of metal nanoparticles sufficiently well.

2.5 Workflow

The workflow for the whole surface atoms labelling pipeline is illustrated by Fig. 2. All written codes and data that support the findings of this study are openly available at **https://github.com/Jon-Ting/metal-nanoparticle-surface-atom-labelling**.

3 Results

3.1 Data visualisation

We first visualise the data distribution for the atoms described in the high dimensional space for a selection of Pd nanoparticles in Fig. 3 using the manifold learning method t-distributed stochastic neighbour embedding. The details of the method and more results are included in ESI† (section S2.2). The points (corresponding to atoms) projected to the

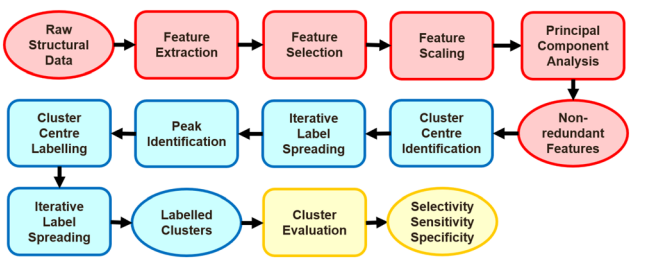

Fig. 2 Workflow for clustering the atoms of metal nanoparticles. The red, blue, and yellow components correspond to data preparation, atom clustering, and result evaluation steps, respectively.

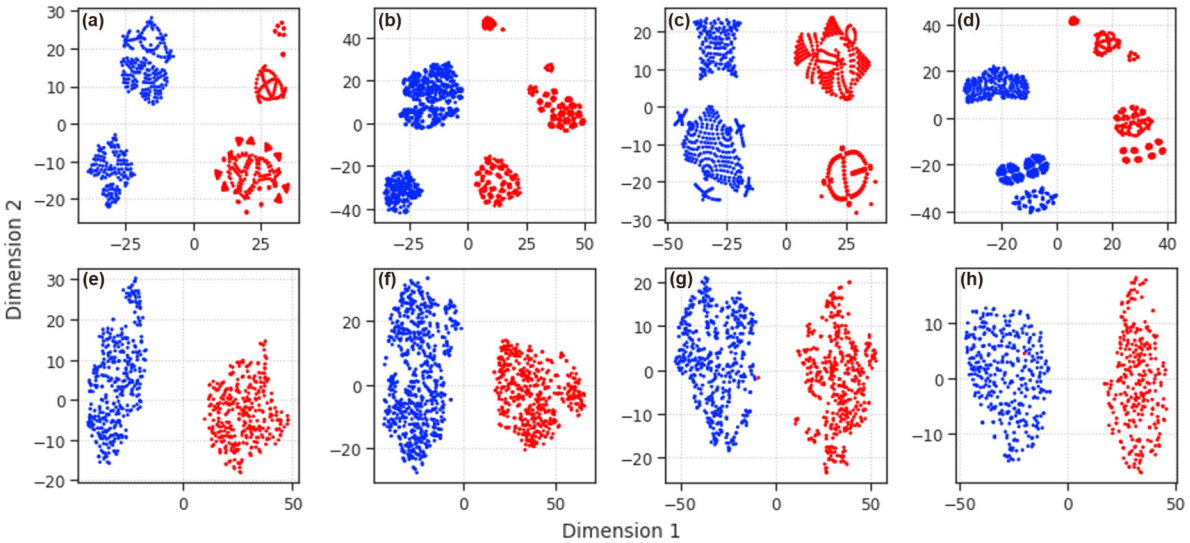

Fig. 3 Mapping of the high dimensional data sets for the (a) octahedron, (b) rhombic dodecahedron, (c) tetrahedron, and (d) truncated octahedron nanoparticles simulated at 0 K, the (e) octahedron, (f) rhombic dodecahedron, and (g) tetrahedron nanoparticles simulated at 523 K, and (h) the smallest disordered nanoparticle simulated at 723 K, onto 2D manifold learnt *via* *t*-distributed stochastic neighbour embedding. The red and blue points correspond to surface and bulk atoms, respectively.

reduced 2D manifold are coloured based on whether the atoms are identified as a bulk or surface atom by NCPac[33] *via* the cone algorithm.[50] The details of the algorithm are provided in ESI† (section S1.6.1). This visualisation shows that the surface and bulk atoms always form distinct groups, so it is reasonable to assume that ILS would also be able to distinguish between groups corresponding to the surface and bulk atoms. We can also see that there are more detailed sub-groups, indicating there are multiple surface (and bulk) clusters with different characteristics. The correlation matrices, and cumulative explained variance plots are included in ESI† (sections S7 and S8).

3.2 Grouping all atoms

Fig. 4 show the R_{min} plots obtained from the first pass of ILS through all atoms of ordered and disordered Pd nanoparticles, coloured by the surface and bulk atoms. It can be seen from the figures that it is straightforward for ILS to distinguish between bulk and surface atoms, particularly the ordered nanoparticles, just by dividing the plot at the central peak. While the bulk atoms seem to be mixed within the surface atoms for some disordered nanoparticles, identifying the central peak and rerunning ILS with a cluster centre on both sides of the peak allows ILS to find the clusters that

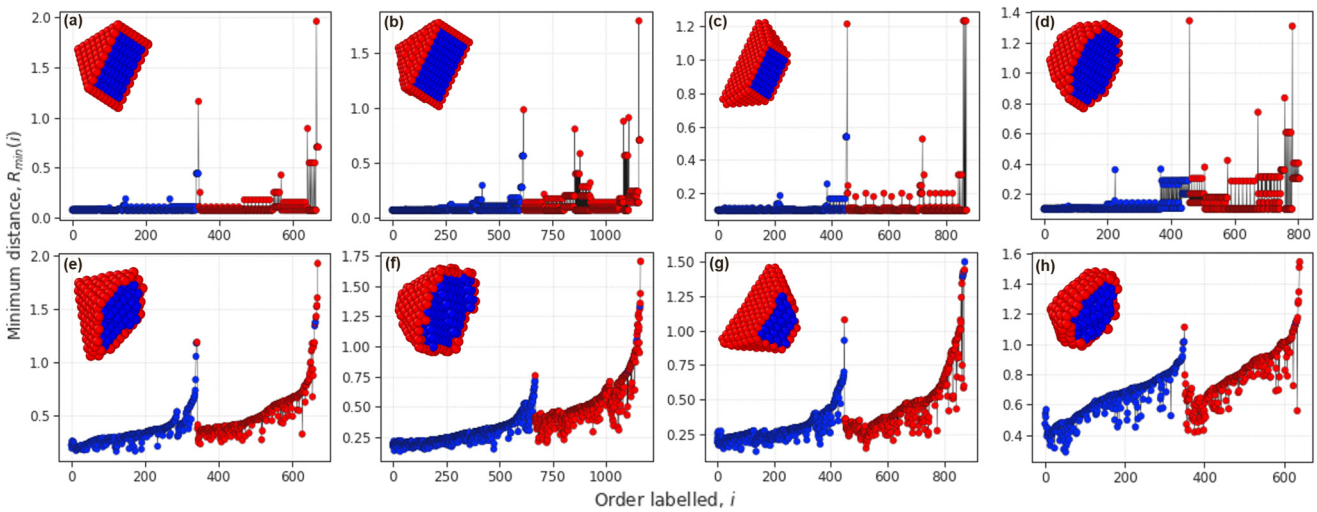

Fig. 4 Minimum distance plots for the (a) octahedron, (b) rhombic dodecahedron, (c) tetrahedron, and (d) truncated octahedron nanoparticles simulated at 0 K, the (e) octahedron, (f) rhombic dodecahedron, and (g) tetrahedron nanoparticles simulated at 523 K, and (h) the smallest disordered nanoparticles simulated at 723 K. The nanoparticles are sliced across the *x*-axis or (100) plane. The points are coloured by whether they are surface (red) or bulk (blue) atoms.

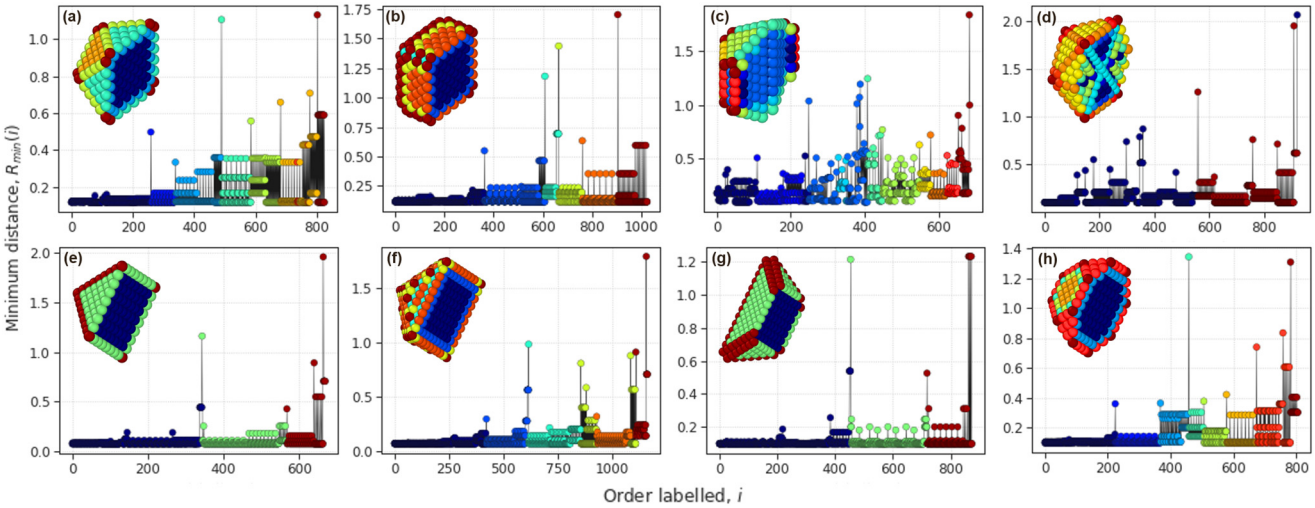

Fig. 5 Minimum distance plots for the (a) cuboctahedron, (b) cube, (c) decahedron, (d) icosahedron, (e) octahedron, (f) rhombic dodecahedron, (g) tetrahedron, and (h) truncated octahedron nanoparticles simulated at 0 K, coloured by clusters identified. The nanoparticles are sliced across the *x*-axis or (100) plane.

correspond well to the surface and bulk atoms. The relevant results are included in ESI† (section S9).

Fig. 5 shows the ILS clusters obtained from clustering all atoms of the ordered nanoparticles, and showed good evidence that the algorithm can be used to identify different types of surface and subsurface layers in nanoparticles, which are crucial for density functional theory studies for catalytic applications.[51,52] The sensitivity of the peak identifying algorithm is tuneable to obtain coarser or more refined details of the subsurface structures, as illustrated in ESI† (section S10). However, it is not trivial to decide what values to tune them to, and this is deferred to be explored in future work. While threshold values could be set based on domain knowledge, to preserve the degree of autonomy of the clustering pipeline, we have used the same threshold values based on the testing on ordered nanoparticles in this work.

3.3 Grouping surface atoms

The final R_{min} plots for the surface atoms of ordered nanoparticles are shown in Fig. 6, with the sample colours corresponding to the identified clusters. This confirms that ILS is capable of identifying the surface patterns deemed important for the catalytic performance of metal nanoparticles, *i.e.* corners, different edges and sub-edges, and different facets.[53,54] The confirmation that the final clustering results have been obtained is achieved by verifying that there are no obvious peaks in the R_{min} plots for each cluster of each nanoparticle. Readers are directed to ESI† (section S11) for more information.

The final R_{min} plots for the surface atoms of disordered nanoparticles, coloured by the identified clusters, are shown in Fig. 7. It was discovered that the features that allow

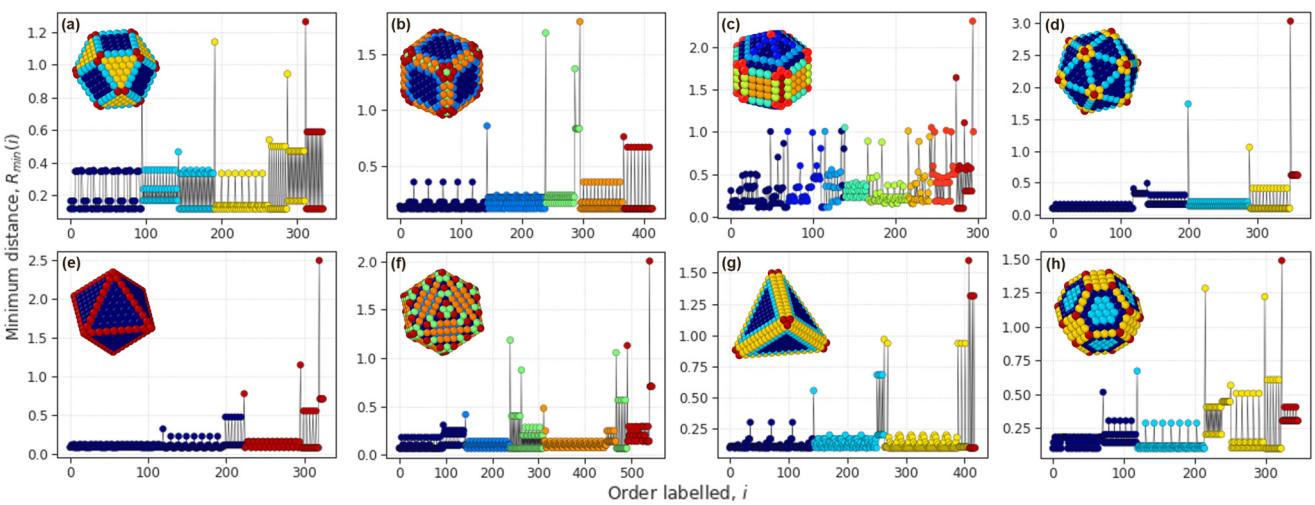

Fig. 6 Iterative label spreading minimum distance plots for the (a) cuboctahedron, (b) cube, (c) decahedron, (d) icosahedron, (e) octahedron, (f) rhombic dodecahedron, (g) tetrahedron, and (h) truncated octahedron nanoparticles simulated at 0 K, coloured by clusters identified.

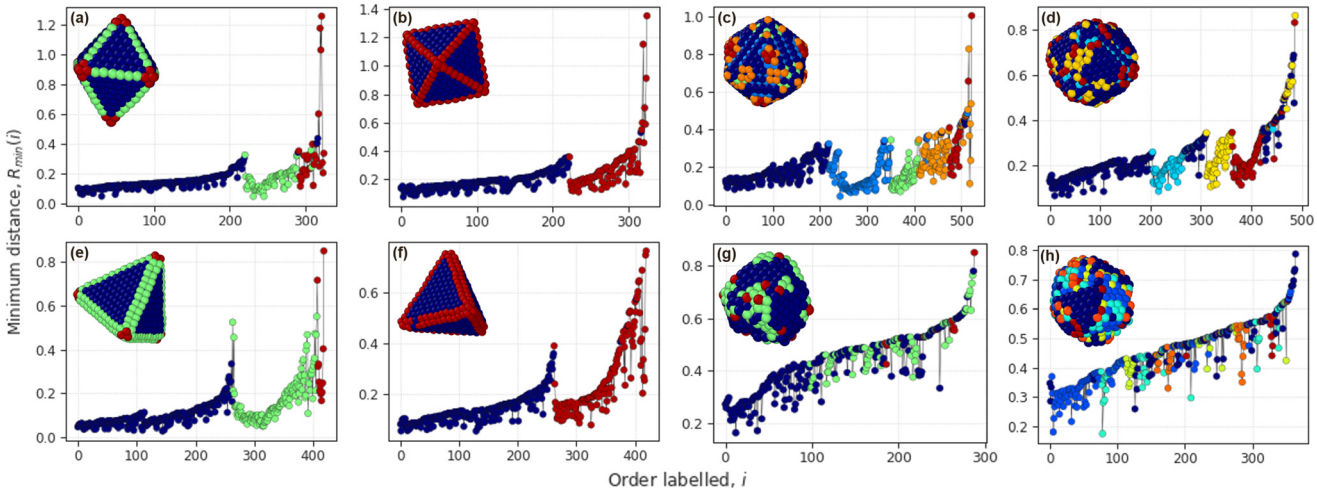

Fig. 7 Iterative label spreading minimum distance plots for octahedron nanoparticles simulated at (a) 323 K and (b) 523 K, rhombic dodecahedron nanoparticles simulated at (c) 323 K and (d) 523 K, tetrahedron nanoparticles simulated at (e) 323 K and (f) 523 K, and (g) small- and (h) medium-sized disordered nanoparticles simulated at 723 K, coloured by clusters identified.

surface characteristics on ordered nanoparticles to be distinguished do not necessarily have the same utility for disordered nanoparticles. Therefore, a smaller set of features (determined from an experiment testing different combinations of descriptors) are used to obtain the results for the disordered nanoparticles. The results obtained using the original (all features) and other feature spaces are included in ESI† (section S12). It is also noted that, as the final clusters are obtained from a second pass of ILS and are projected back onto the R_{min} plots obtained from the first pass of ILS, the colours do not necessarily appear to be consecutive in the plot. The patterns identified confirm that the algorithm is able to group atoms with similar surface patterns that are difficult to describe with any single catalytic variable, and provides good evidence that the algorithm is capable of combining the information in the high dimensional feature space, which will be important when labelling disordered nanoparticles. Further illustrations on the ability of the algorithm to identify the peaks where human eyes might fail are included in ESI† (Section S13).

3.4 Surface pattern analysis

The feature profiles of the surface patterns identified can be analysed to inform researchers about the characteristics of any surface atom groups of interest, which will be determined *via* the cluster evaluation step in this work. As an example, the box plots of a selected set of features of the surface clusters identified for an ordered CO nanoparticle simulated at 0 K, and a disordered nanoparticle simulated at 723 K are shown in Fig. 8 (the box plots for the other features are included in section S14 of ESI†), with the corresponding clusters visualised in Fig. 9. The four clusters for the ordered nanoparticle are composed of atoms with very different

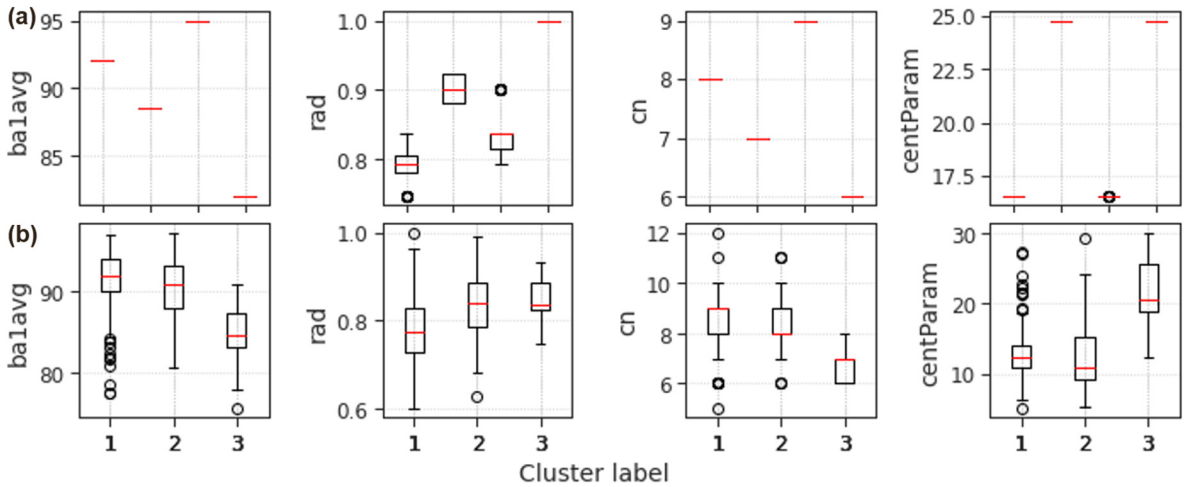

Fig. 8 Box plots of a selected set of features for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K, with the medians marked by red lines. The explanation of the features are provided in ESI† (section S1.5).

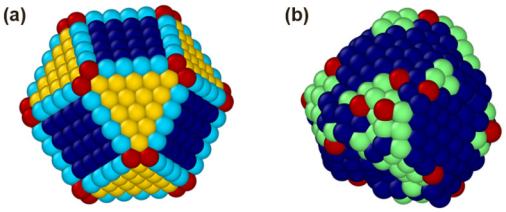

Fig. 9 Identified surface clusters for the (a) ordered cuboctahedron nanoparticle simulated at 0 K and (b) smallest disordered nanoparticle simulated at 723 K. The cluster identifier number ascends from blue (smallest number) to red (biggest number). Clusters 1, 2, 3, and 4 of (a) thus correspond to the {100} facets, edges, {111} facets, and corners, respectively.

coordination numbers (which is correlated with the average bond angles), radial distance from nanoparticle centre, and certain Steinhardt parameters. Fig. 9(a) reveals that clusters 1, 2, 3, and 4 correspond to the atoms on {100} facets, edges, {111} facets, and corners, respectively. However, even without reference to the figure, we can infer this from these feature profiles alone. For example, the surface atoms in cluster 4 are the corner atoms as they are furthest away from the centre of the nanoparticle, bonded to a lower number of neighbours, and have a lower symmetry.

The differences between the three surface patterns identified in the disordered nanoparticle are more subtle. Nonetheless, it can be concluded that atoms in cluster 1 tend to be more deeply embedded among the other surface atoms (as indicated by its relatively lower median normalised radial distance from the nanoparticle centre and higher coordination number compared to other clusters) and are more symmetric locally. Fig. 9(b) indicates that they are the most facet-like atoms. The atoms in cluster 2 are very far from the nanoparticle centre (protruded from the facet-like atoms), have coordination values that fall in between the other clusters, and are rather symmetric locally. We refer to them as the most step-like atoms based on the visualisation in Fig. 9(b). Cluster 3 atoms have the lowest coordination and shortest average distance to the neighbouring atoms. Fig. 9(b) shows that they are the most adatom-like atoms (the opposite of vacancies) on the surface. The differences in all features collectively make the surface pattern of each cluster unique, in a way that is difficult to be described by any single geometric variable. We also note that all of our features are structural in this work. The insight obtained could be even more informative for catalysis if electronic factors are included in the feature space.

3.5 Catalytic evaluation of surface patterns

In this section, we compare the evaluation results between the two nanoparticles investigated in section 3.4, in terms of the selectivity, specificity, and sensitivity of the surface clusters identified toward some selected chemical reactions, namely ORR, COOR, and RCORRR. The scores of all clusters of the surface patterns identified for both nanoparticles toward the reactions are listed in Table 2. The internal evaluation scores of the clustering results for all nanoparticles and the comparisons between the catalytic weighting profiles and the surface cluster GCN distributions for the two nanoparticles are included in ESI† (sections S15 and S16, respectively).

Table 2 Selectivity (SEL), specificity (SPC), and sensitivity (SEN) toward oxygen reduction reaction (ORR), carbon monoxide oxidation reaction (COOR), and aliphatic ketone reduction reaction (RCORRR) for all surface clusters of an ordered cuboctahedron nanoparticle simulated at 0 K and a disordered nanoparticle simulated at 723 K. All three metrics are bound within the range of [0, 1]

Nanoparticle	Reaction	Cluster	SEL	SPC	SEN
Ordered	ORR	1	0.872	0.987	0.093
		2	0.764	0.000	0.000
		3	0.918	0.983	0.236
		4	0.687	0.000	0.000
	COOR	1	0.909	0.985	0.211
		2	0.984	0.981	0.678
		3	0.863	0.980	0.149
		4	0.907	0.164	0.190
	RCORRR	1	0.952	0.993	0.183
		2	0.940	0.991	0.399
		3	0.906	0.990	0.230
		4	0.863	0.859	0.237
Disordered	ORR	1	0.900	0.166	0.207
		2	0.813	0.038	0.011
		3	0.737	0.000	0.000
	COOR	1	0.881	0.773	0.316
		2	0.967	0.851	0.478
		3	0.957	0.866	0.430
	RCORRR	1	0.924	0.891	0.377
		2	0.989	0.926	0.436
		3	0.913	0.936	0.303

It is observed that cluster 3 ({111} surfaces) of the cuboctahedron nanoparticle has the highest scores toward ORR. This is in accordance with the findings in the past, where only sites with the same number of first-nearest neighbours as {111} terraces but with increased number of second nearest-neighbours are predicted to have superior catalytic activity over the atoms on the {111} terraces.[17] The surface pattern that is expected to best contribute to the catalytic activity toward COOR in terms of selectivity, specificity, and sensitivity is found to be cluster 2, which corresponds to the edge atoms. This once again agrees with the findings in the previous work, where the maximal activity is reached on the step edges of the electrodes, which has GCN of approximately 5.4.[23] The most catalytically active sites for RCORRR were determined to be the steps on {110} facets for 2-propanol production and the {110} steps on {510} facets for propane production.[22] The cluster on the cuboctahedron nanoparticle surface with the closest surface patterns to these surface structures is cluster 2 (the edge atoms). While cluster 1 (the {100} facets) has slightly higher selectivity and specificity scores, cluster 2 is deemed to be superior overall when sensitivity is taken into account. The finding that no adsorption nor hydrogenation occur at the {100} and {111} facets of Pt electrode[22] is also reflected by the relatively lower

sensitivity of clusters 1 and 3 toward RCORRR. These matching observations with the findings in the literature validates the reliability of the algorithm in evaluating the clusters according to the catalytic relevance to different chemical reactions, and builds our confidence for the patterns discovered by the algorithm among the disordered nanoparticles.

The work of Rossi indicated that the most active sites for ORR tend to be the more deeply embedded atoms on relatively flat surface, with GCN values within the range of [7.5, 8.3].[8,18] This is in agreement with the prediction in this work, where cluster 1 atoms (which are the most deeply embedded and facet-like) are predicted to exhibit superior catalytic performance (in terms of all three evaluation metrics) for ORR, and cluster 3 atoms (which are the most adatom-like) are predicted to be largely irrelevant (with 0 specificity and sensitivity) for ORR. Jørgensen and Grönbeck found that the edges and corners of Pt nanoparticles tend to dominate the catalysis of COOR, which is only facilitated by the facets when the edges and corners are poisoned.[24] Specifically, the edges are more active than the corners.[25] This study also predicts that the step-like and adatom-like atoms in clusters 2 and 3 have higher catalytic performance for COOR than the facet-like cluster 1 atoms. The trend of activity is also similar to the literature,[25] with cluster 2 atoms having higher selectivity and sensitivity but slightly lower specificity than cluster 3. For RCORRR, the performance of the most step-like cluster 2 atoms is predicted to be superior over the other clusters, in accordance to the findings of Bondue *et al.*[22]

While more research into these surface patterns using electronic structure methods would be beneficial to confirm this prediction, and identify underlying mechanisms, the present work demonstrates that these patterns can be found and labelled automatically.

4 Conclusions

We have presented a new data-driven approach to identify the patterns among the surface atoms of simulated metal nanoparticles and characterise their catalytic potentials. The surface atoms are grouped into patterns based on their similarities in the high dimensional feature space using the iterative label spreading clustering algorithm. After atomistic features are extracted and processed, a complete workflow pipeline is demonstrated on a data set of simulated Pd nanoparticles. This approach can be generalised to other atomistic objects, and automated to label entire molecular dynamics trajectories.

The possibility of using the surface atom clusters as the performance indicators of the catalytic potential of the nanoparticles was investigated. The surface patterns were found to provide a reliable, purely unsupervised labelling scheme for nanoparticle surface atoms, capable of identifying complicated surface patterns that may be unintuitive to researchers, but highly relevant to different catalytic reactions. This approach is significantly faster than electronic structure simulations, capable of characterising large nanoparticles, and could replace current, simplistic labelling schemes that fail to capture multi-atom effects. The surface patterns can act as a surrogate label for their catalytic activities by allowing the catalytic contribution of the surface pattern groups of any simulated nanoparticle toward chemical reactions to be quantified. This pipeline is general and can be automated to remove the labelling bottleneck that prevents more widespread usage of large data sets and molecular dynamics simulation trajectories in the study of nanocatalysts.

It was found that the features used to distinguish the surface characteristics of ordered nanoparticles may not be as effective for disordered nanoparticle structures. Consequently, as the degree of disorder in the nanoparticle structures increases, the original feature space gradually loses its capability to capture the surface patterns. The surface patterns for clusters for disordered nanoparticles can be detected by (i) refining the selection of features for disordered nanoparticles, and/or (ii) tuning the sensitivity of the peak identification algorithm. However, it is not trivial to decide the optimal groups of features and optimal values for the tuning, hence dedicated future work is planned to investigate these issues.

We also note that the evaluation metrics in this work are based on the activity maps for platinum and copper nanoparticles. However, the transferability of the maps across different metals is unknown. Ideally, the nanoparticles should be evaluated by the catalytic weightings obtained from the activity maps computed using the nanoparticles of the same type of elements. Nonetheless, the methodology developed here will be applicable as soon as such maps are available from the literature.

Data availability

This study was carried out using publicly available palladium nanoparticle data from CSIRO Data Access Portal at **https://doi.org/10.25919/epxd-8p61**. Data generated for this article, including the extracted features, code scripts, and figures are available at **https://github.com/Jon-Ting/metal-nanoparticle-surface-atom-labelling**. The original source codes for NCPac (version 1) and SYMMOL (version October 22nd 2002) can be found at **https://doi.org/10.25919/tfv3-he58** and **https://github.com/fxcoudert/symmol**, respectively.

Author contributions

Jonathan Yik Chang Ting: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing – original draft, writing – review and editing. George Opletal: software, writing – review and editing. Amanda S. Barnard: conceptualisation, resources, funding acquisition, writing – review and editing, supervision.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Computational resources for this project were supplied by the National Computing Infrastructure (NCI) [grant numbers p00 and q27]. Jonathan Yik Chang Ting is grateful for the support from Australian National University *via* the University Research Scholarship. The authors are grateful to Dr. Federico Calle-Vallejo for providing the equations for the activity map of the aliphatic ketone reduction reaction.

References

1 A. T. Bell, *Science*, 2003, **299**, 1688–1691.
2 T. S. Rodrigues, A. G. M. da Silva and P. H. C. Camargo, *J. Mater. Chem. A*, 2019, **7**, 5857–5874.
3 N. Narayan, A. Meiyazhagan and R. Vajtai, *Materials*, 2019, **12**, 3602.
4 C.-H. Cui and S.-H. Yu, *Acc. Chem. Res.*, 2013, **46**, 1427–1437.
5 S. Schauermann and H.-J. Freund, *Acc. Chem. Res.*, 2015, **48**, 2775–2782.
6 C. Vogt and B. M. Weckhuysen, *Nat. Rev. Chem.*, 2022, **6**, 89–111.
7 K. Rossi, G. G. Asara and F. Baletto, *Phys. Chem. Chem. Phys.*, 2019, **21**, 4888–4898.
8 K. Rossi, G. G. Asara and F. Baletto, *ACS Catal.*, 2020, **10**, 3911–3920.
9 E. Gazzarrini, K. Rossi and F. Baletto, *Nanoscale*, 2021, **13**, 5857–5867.
10 R. M. Jones, K. Rossi, C. Zeni, M. Vanzan, I. Vasiljevic, A. Santana-Bonilla and F. Baletto, *Faraday Discuss.*, 2023, **242**, 326–352.
11 X. Ma and H. Xin, *Phys. Rev. Lett.*, 2017, **118**, 036101.
12 M. J. Piotrowski, P. Piquini and J. L. D. Silva, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2010, **81**, 155446.
13 F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet and D. Loffreda, *Angew. Chem., Int. Ed.*, 2014, **53**, 8316–8319.
14 F. Calle-Vallejo and A. S. Bandarenka, *ChemSusChem*, 2018, **11**, 1824–1828.
15 F. Calle-Vallejo, *Adv. Sci.*, 2023, **10**, 2207644.
16 J. Kari, J. P. Olsen, K. Jensen, S. F. Badino, K. B. Krogh, K. Borch and P. Westh, *ACS Catal.*, 2018, **8**, 11966–11972.
17 F. Calle-Vallejo, J. Tymoczko, V. Colic, Q. H. Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann and A. S. Bandarenka, *Science*, 2015, **350**, 185–189.
18 M. Rück, A. Bandarenka, F. Calle-Vallejo and A. Gagliardi, *J. Phys. Chem. Lett.*, 2018, **9**, 4463–4468.
19 M. Núñez, J. L. Lansford and D. G. Vlachos, *Nat. Chem.*, 2019, **11**, 449–456.
20 Z. Zhao, Z. Chen, X. Zhang and G. Lu, *J. Phys. Chem. C*, 2016, **120**, 28125–28130.
21 L. G. Verga, P. C. Mendes, V. K. Ocampo-Restrepo and J. L. D. Silva, *Catal. Sci. Technol.*, 2022, **12**, 869–879.
22 C. J. Bondue, F. Calle-Vallejo, M. C. Figueiredo and M. T. M. Koper, *Nat. Catal.*, 2019, **2**, 243–250.
23 F. Calle-Vallejo, M. D. Pohl and A. S. Bandarenka, *ACS Catal.*, 2017, **7**, 4355–4359.
24 M. Jørgensen and H. Grönbeck, *ACS Catal.*, 2017, **7**, 5054–5061.
25 M. Jørgensen and H. Grönbeck, *Angew. Chem., Int. Ed.*, 2018, **57**, 5086–5089.
26 M. Hu, J. He, R. Guo, W. Yuan, W. Xi, J. Luo and Y. Ding, *Catal. Commun.*, 2020, **146**, 106129.
27 C. Roncaglia and R. Ferrando, *J. Chem. Inf. Model.*, 2023, **63**, 459–473.
28 C. Zeni, K. Rossi, T. Pavloudis, J. Kioseoglou, S. de Gironcoli, R. E. Palmer and F. Baletto, *Nat. Commun.*, 2021, **12**, 6056.
29 C. Zeni, K. Rossi, A. Glielmo and S. de Gironcoli, *J. Chem. Phys.*, 2021, **154**, 224112.
30 R. Drautz, *Phys. Rev. B*, 2019, **99**, 14104.
31 A. Barnard and G. Opletal, *Palladium Nanoparticle Data Set. v1*, 2019, **https://data.csiro.au/collection/csiro:40618**.
32 X. W. Zhou, R. A. Johnson and H. N. G. Wadley, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2004, **69**, 144113.
33 G. Opletal, J. Y. C. Ting and A. S. Barnard, *NCPac*, 2024, **https://doi.org/10.25919/tfv3-he58**.
34 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.
35 S. Menon, G. D. Leines and J. Rogal, *J. Open Source Softw.*, 2019, **4**, 1824.
36 T. Pilati and A. Forni, *J. Appl. Crystallogr.*, 1998, **31**, 503–504.
37 A. J. Parker and A. S. Barnard, *Adv. Theory Simul.*, 2019, **2**, 1–8.
38 S. Lloyd, *IEEE Trans. Inf. Theory*, 1982, **28**, 129–137.
39 M. Ester, H.-P. Kriegel, J. Sander and X. Xu, *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.
40 F. Murtagh and P. Contreras, *WIREs Data. Mining. Knowl. Discov.*, 2012, **2**, 86–97.
41 A. J. Parker and A. S. Barnard, *Nanoscale Horiz.*, 2020, **5**, 1394–1399.
42 A. J. Parker, G. Opletal and A. S. Barnard, *J. Appl. Phys.*, 2020, **128**, 1–11.
43 A. J. Parker and A. S. Barnard, *Nanoscale Horiz.*, 2021, **6**, 277–282.
44 P. J. Rousseeuw, *J. Comput. Appl. Math.*, 1987, **20**, 53–65.
45 T. Caliński and J. Harabasz, *Commun. Stat.*, 1974, **3**, 1–27.
46 D. L. Davies and D. W. Bouldin, *IEEE Trans. Pattern Anal. Mach.*, 1979, **PAMI-1**, 224–227.
47 J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 937–943.

48 G. Carchini, N. Almora-Barrios, G. Revilla-López, L. Bellarosa, R. García-Muelas, M. García-Melchor, S. Pogodin, P. Błoński and N. López, *Top. Catal.*, 2013, **56**, 1262–1272.
49 S. D. Miller, N. Inoğlu and J. R. Kitchin, *J. Chem. Phys.*, 2011, **134**, 104709.
50 Y. Wang, S. Teitel and C. Dellago, *J. Chem. Phys.*, 2005, **122**, 214722.
51 P. Tiruppathi, J. J. Low, A. S. Chan, S. R. Bare and R. J. Meyer, *Catal. Today*, 2011, **165**, 106–111.
52 M. Gulumian, C. Andraos, A. Afantitis, T. Puzyn and N. J. Coville, *Int. J. Mol. Sci.*, 2021, **22**, 8347.
53 C. Burda, X. Chen, R. Narayanan and M. A. El-Sayed, *Chem. Rev.*, 2005, **105**, 1025–1102.
54 Y. Li and G. A. Somorjai, *Nano Lett.*, 2010, **10**, 2289–2295.