# Classification of Platinum Nanoparticle Catalysts using Machine Learning

A. J. Parker,[1, a)] G. Opletal,[1, b)] and A. S. Barnard[2, c)]
[1)]*CSIRO Data61, Docklands VIC 3008, Australia*
[2)]*ANU Research School of Computer Science, Acton ACT 2601, Australia*

(Dated: 21 May 2020)

Computer simulations and machine learning provide complementary ways of identifying structure/property relationships that are typically targeting toward predicting the ideal singular structure to maximise the performance on a given application. This can be inconsistent with experimental observations that measure the collective properties of entire samples of structures that contain distributions or mixture of structures, even when synthesized and processed with care. Metallic nanoparticle catalysts are an important example. In this study we have used a multi-stage machine learning workflow to identify the correct structure/property relationships of Pt nanoparticles relevant to oxygen reduction (ORR), hydrogen oxidation (HOR) and hydrogen evolution (HER) reactions. By including classification prior to regression we identified two distinct classes of nanoparticles, and subsequently generate the class-specific models based on experimentally relevant criteria that are consistent with observations. These multi-structure/multi-property relationships, predicting properties averaged over a large sample of structures, provide a more accessible way to transfer data-driven predictions into the lab.

## I. INTRODUCTION

While it has been well established that the size, shape and surface structure of metallic nanoparticles are responsible for their performance in a variety of applications, complete control over the structure remains challenging[1–4] due to competition and collaboration between growth kinetics and thermodynamics during synthesis.[5–9] Considerable effort has been directed toward controlling the size and shape of metal nanoparticles,[10–15] but many samples persistently contain imperfect shapes, disordered lattices, and defective surfaces.[16–18]

In the case of platinum, it is known that nanoparticles with controlled sizes and shapes, characterized by surface facets and in specific crystallographic orientations, can be used to tune the sensitivity and selectivity of many important catalytic reactions.[19,20] Important factors contributing to the morphology of individual nanoparticles include the type and concentration of the precursor, the reducing agent and stabilizer, the introduction of seeds or foreign species,[21] the impact of twinning and structural defects,[22,23] and temperature. Among the methods developed to control these factors, solution-phase synthesis is highly versatile[24–27] and uses the reduction and decomposition of a metal precursor in the presence of a surfactant to engineer the structure of platinum.[28–31] Understanding the relationship between these structural and processing parameters and the desirable properties is one of the goals of rational nanoparticle design, particularly in the engineering of nanocatalysts, and so the extensive body of experimental literature has been augmented with theoretical and computational studies that provide insight into the properties of specific structures. For example, a detailed computational screening of surface structures for new nanocatalysts has been performed for the methanation reaction,[32] but due to the high computational cost of the electronic structure calculations was limited to only a few dozen instances.[33–37] Studies such as these providing important information on reaction efficiency, but are vulnerable to selection bias, confirmation bias and reporting bias,[38] as each of the limited systems being studied was carefully pre-determined. They also fail to capture the averaged efficiency of real samples that contain a distribution or mixture of sizes, shapes and defects, and are not in the ground state.

The mismatch between the averaged performance measured during experimental studies and the specific (and limited) focus of conventional computational studies presents a problem when attempts are made to translate computational predictions into the lab. This problem is not unique to platinum nanoparticles, but is particularly relevant given the strong connection between the highly complex surface structure and catalytic performance. Machine learning (ML) methods, however, are ideally suited to studying the complex multi-structure correlations that are difficult to identify using conventional computational methods, and are free from some of the assumptions and biases introduced by human researchers. It has been previously established that ML can produce parametric functions of structural features capable of accurate predictions of useful properties based on a large set of atomistic simulations.[39] By combining a sufficiently large and diverse ensemble of candidate nanostructures generated using conventional simulations with an appropriate regressor it is possible to identify the set of features that drive performance,[40] and in some cases conditions required to deliver the right structures in practice.[41] ML is also capable of determining classes of like-structures based on similarity, and then correlate these classes with some performance indicators to provide a more averaged response to structure/property prediction, akin to measuring a diverse mix of sizes and shapes.[42,43] Most importantly, ML is providing to be invaluable in the modern design of catalysts.[44,45]

In this study, we use ML to predict the different classes of platinum nanocatalysts based on structural features and two widely used synthetic processing conditions, and identify class-specific structure/property relationships to established

[a)] Electronic mail: amanda.parker@data61.csiro.au
[b)] Electronic mail: george.opletal@data61.csiro.au
[c)] Electronic mail: amanda.s.barnard@anu.edu.au

indicators of efficient hydrogen evolution reactions (HER), hydrogen oxidation reactions (HOR) and oxygen reduction reactions (ORR).[19,20,46–48] We have used an ensemble of 1300 unique platinum nanoparticles generated from molecular dynamics (MD) trajectories that sample a large variety of different temperatures and growth rates, and apply sophisticated clustering, classification and regression algorithms to identify how the overall characteristics of each class may enhance or suppress performance. As we will show, the classification of the nanoparticles into ordered and disordered structures is an important first step to predicting the correct structure/property relationships with machine learning. Following the separation of the particles into these classes, regression is able to accurately identify the important structural features responsible for ORR and HER/HOR reactions in agreement with experimental observation, but without prior classification, the predictions are confused and provide no clear path to impact.

## II. DATA SET AND METHODS

In this study, we have used an existing set of atomistic platinum nanoparticles originally generated using molecular dynamics to simulate growth *via* random addition of single Pt atoms and unguided sintering and coalescence events, or with experimentally relevant morphologies relaxed using molecular dynamics at elevated temperatures to allow for the formation or annealing of twins. The set contains particles grown using different atomic deposition rates ($tau$) and temperatures ($T$), to capture the effects of inhomogeneous reaction kinetics and thermal fluctuations within a range of values characteristic of experiments. The set is available for download, with detailed information on the simulation procedure and an extensive list of 179 structural features based on atomic, crystallographic and topological descriptors.[49] This list was reduced to 121 dimensions (including structural and processing features) by eliminating features with zero variance, and then appropriately normalized so that all features occupy the range 0 to 1. This data set supersedes previous versions[50] has it is larger, includes ordered shapes as well as anisotropic and disordered particles, has a larger range of processing conditions, and a greater number of structural features. Growth time was not included as a feature in this case, as some of the relaxed nanoparticles were included as pre-grown structures.

Three indicators of molar catalytic activity have been used as the target labels. The molar catalytic activity was estimated using a surface coordination number (SCN) scheme that groups types of surface imperfections based on the degree of under-coordination of each surface atom, and the similarity with respect to known surface features that have been shown to enhance different catalytic reactions.[19,20,46–48] Under this scheme *Surface Defects* include all adatoms in configurations (“top", “bridge" and “hollow") where the Pt-coordination number can be 1, 2 or 3; *Surface Microstructures* include surface “kinks" and “steps", where the Pt coordination number can be 4, 5, 6 or 7; and *Surface Facets* include configurations (in any *hkl* orientation) where the Pt coordination number can be 8, 9, 10 or 11 (recalling the coordination number of Pt atom in the bulk is 12). Although these assignments may seem ambiguous, each of these groups are linked to a specific catalytic reaction, and were originally determined based on a full survey of the literature.[51–60] For example, *Surface Facet*-driven catalytic activity is suitable for hydrogenation reactions, whilst nanoparticles with *Surface Microstructure*-driven activity are more efficient to catalyse combustion reactions. A theoretical hydrogenation/combustion selectivity can be defined as the ratio between *Surface Facet*-driven catalytic activity and *Surface Microstructure*-driven catalytic activity. This scheme has been shown to be suitable for investigating active sites on nanoparticle surface in the past,[61–63] and has been successfully combined with theoretical analysis[64] and machine learning.[65]

In the present study, we concentrate on *Surface Microstructures* and *Surface Facets*, as we seek insights into HER, HOR and ORR.

### A. Clustering

Clustering methods are unsupervised pattern recognition techniques that group samples based on a similarity index, without reference to target labels. There are many different clustering methods available, each with advantages and disadvantages.[66] In this study we have used a new clustering method that has the advantage of including hyper-parameter optimization.[67] Iterative label spreading (ILS) is based on a general definition of a cluster and the quality of a clustering result, and is capable of predicting the number and type of clusters and outliers in advance of clustering, regardless of the complexity of the distribution of the data. ILS can be used to evaluate the results from other clustering algorithms, or perform clustering directly. It has been shown to be more reliable than alternative approaches for simple and challenging cases (such as the null and chain cases) and to be ideal for studying noisy data with high dimensionality and high variance, as is typical for nanoparticle systems.

Direct clustering is achieved using this algorithm by initializing one labeled point and applying ILS to obtain the ordered minimum distance ($R_{min}$(i)) plot, as described in detail in Ref. 67. The number of clusters can be automatically extracted by identifying peaks in the $R_{min}$(i) plot (due to density drops between clusters) that divide the plot into $n$ regions. This can be automated using a continuous wavelet transform peak finding algorithm with smoothing over $p$ points. The smoothing essentially sets the minimum cluster size to identify clusters of no smaller than $p$. One point can be relabelled in each region (preferably at the minima) to run ILS again, and obtain a fully labeled data set with $n$ clusters defined. ILS can also be applied to each individual cluster to confirm that each region is a single cluster that should not be divided further.

### B. Classifiers

Classification is a type of supervised learning where the target labels are also provided with the features. A classifier is trained (using input training data) to recognise how unseen instances relate to some known classes of instances and assigns them accordingly. There are numerous classification algorithms available, and the superiority of one over another depends on the application and the data set.

In this study we have used the (non-linear, non-parametric) Extra Trees Classifier (ETC), which fits a number of randomized decision trees to the training sub-set, and averages over the results to improve the predictive accuracy and control over-fitting. ETC is generally faster than similar estimators and performs well in the presence of noisy features typical of nanoparticle data sets. The hyper-parameters of the ETC were optimised using a grid search (criterion='gini', max_depth=None, max_features='sqrt', min_impurity_split= 0.01, min_weight_fraction_leaf=0, n_estimators=50, class_weight=None, oob_score=False, warm_start=False) and applied using 10-fold cross validation, and a 25/75 test/train split.

Decision trees are trained by recursively splitting the data, but are prone to over-fitting. For this reason, we calculated the learning curve test for convergence of the training and cross validation scores. These results compare well with alternative classifiers, Logistic Regression (linear) and Random Forest (non-linear), which confirmed the results (see Supporting Information).

### C. Regressors

Regression is a type of supervised learning to predict the relationship between the features and a target label. A regressor is trained (using input training data) to recognise a continuous relationship and predict the expected target property for unseen data based on the known features. Just as for classifiers, there are numerous regression algorithms available, and the superiority of one over another depends on the application and the data set.

In this study, we have used the (non-linear, non-parametric) Extra Trees Regressor (ETR) which, in a similar way to the ETC, fits a number of randomized decision trees to a sub-set, and averages over the results. Following classification, regression was performed on each class individually for each target property label. The hyper-parameters were optimised for each class using a grid search (as described later on) and applied using 10-fold cross validation, and a 25/75 test/train split.

The results were compared with the ridge regression (linear) and random forest progression (non-linear), which resulted in more significant under-fitting and over-fitting, respectively (see Supporting Information).

## III. RESULTS

To better understand the average performance of similar types of platinum nanoparticles, clustering and classification were undertaken before regression, in order to identify class-dependent structure/property relationships.

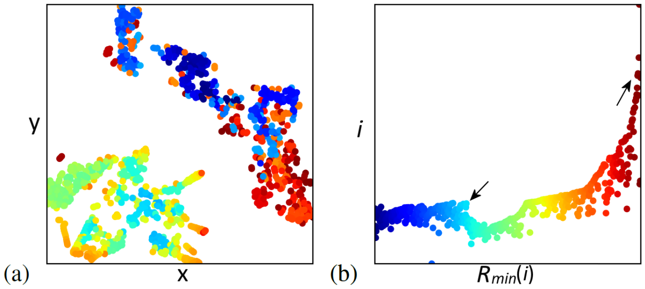

FIG. 1. (a) x-y distribution of the 1300 platinum nanoparticles using t-SNE, based on their similarity in 121 dimensions, and (b) the order-labelled $R_{min}$ plot generated using ILS clustering showing two peaks identifying the two distinct clusters (indicated by the peaks highlighted with arrows). Both plots are colored by the order in which the labels were iterated, from blue to red.

### A. Classification

Using ILS we identified two well defined clusters in the platinum nanoparticle ensemble. Shown in Fig. 1)(a) is the distribution of the set visualised using t-distributed stochastic neighbour embedding (t-SNE)[69], which has successfully been useful in visualising multi-dimensional nanoparticle data sets in the past.[70] In Fig. 1(b) we show the ILS $R_{min}$(i) plot which shows two distinct peaks identifying the two clusters. In both cases, the points have been colored by the order in which the labels were iterated (blue to red) by ILS, and we can see that that dark red points (labeled last), are a greater distance from the prior points, suggesting they may be outliers. This is also supported by the t-SNE distribution where these points appear at the edges of the clusters.

Based on this result we can assign each nanoparticle to a cluster (1 and 2), as shown in Fig. 2(a) and perform classification to determine if the clusters constitute classes. We first removed outliers (reducing the number of instances to 1279), which would otherwise introduce bias into our machine learning models, and applied the ETC. The results are captured in the confusion matrix shown in Fig. 2(b) where there is perfect accuracy, precision and recall. The impact of outlier removal is shown in the Supporting Information. The classes are perfectly separable, based largely on the growth rate ($tau$), various order parameters (q6q6_X), and the growth temperature ($T$). This is shown in the feature importance plot (see Fig. 2(c)), which also indicates that the order parameters based on surface coordination q6q6_S0 (number of surface atoms having 0 nearest neighbours with similar bonding environments as itself), coordination q6q6_S2 (number of surface atoms having 2 nearest neighbours with similar bonding environments as itself), and q6q6_avg_bulk (the average number of bulk atoms having an environment similar to itself) are important. The order parameters are indicative of highly disordered surfaces where few atoms are surrounded by similarly coordinated atoms. These are followed by $T$, and a number of other order parameters similarly indicative of neighbours with low structural similarity (both surface and total number of atoms, which are the same when the coordination numbers are small).

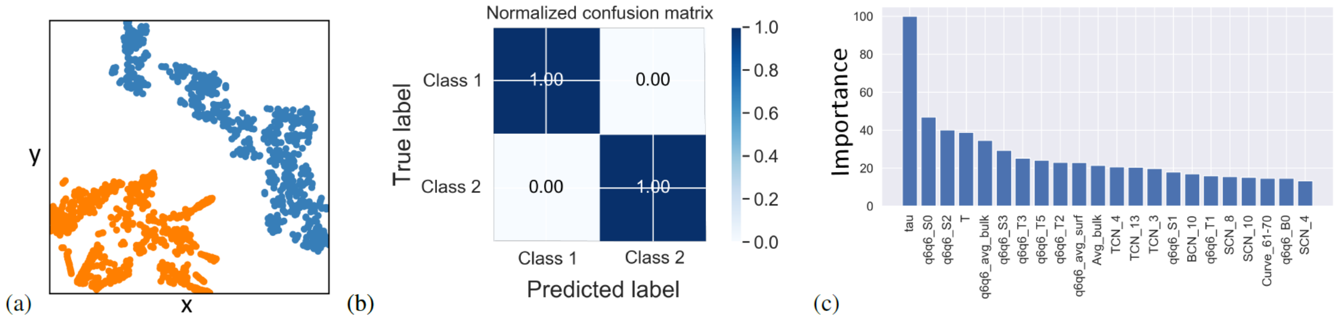

FIG. 2. (a) Distribution of the 1300 platinum nanoparticles using t-SNE, colored by the cluster assigned using ILS, (b) the confusion matrix showing the classes are perfectly separable, and (c) the feature importance histogram showing the classes are largely determined by the processing conditions, *tau* and $T$, the order parameters based on surface coordination q6q6_S0 (number of surface atoms having 0 nearest neighbours with similar bonding environments as itself), coordination q6q6_S2 (number of surface atoms having 2 nearest neighbours with similar bonding environments as itself), and q6q6_avg_bulk (the average number of bulk atoms having an neighbours with bonding environment similar to itself); this is followed by a number of other order parameters indicative of neighbours with low structural similarity.

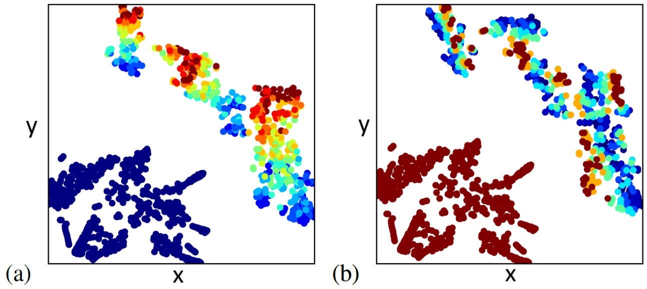

FIG. 3. t-SNE distribution of set colored by the normalized processing features, (a) the relative growth rate, *tau*, (b) the relative growth temperature, $T$. Both plots are colored from 0 to 1 (from blue to red).

This Fig. shows only the top 21 features of the 121 features used to train the model. Fig. 3 shows the distribution of these two important features across the set, where we can see that the Class 1 (upper right) exhibits a complicated spread of $T$, but obvious trends in *tau*. Class 2 (lower left) contains "as-grown" platinum nanoparticles seeds relaxed using molecular dynamics at 673 K.

Based on this classification we next examined the distribution of the target property labels (the catalytic indicators, *Surface Microstructures* and *Surface Facets*) in each of the clusters, as shown in Figs. 4(b) and 4(d), respectively. Here we can see that both clusters include the entire range of *Surface Microstructure* and *Surface Facet* concentrations. This is confirmed when we color encode the t-SNE plots with these property labels, where we can see that the distribution for each of these properties is different for the two classes (which were only trained on the structural features), suggesting that they will have different structure/property relationships, ordered particles will behave very differently to disordered particles. A comparison of the Figs. 4(a) and 4(c) with Figs. 3(a), 3(b) and 3(c) suggest a stronger relationship between these property indicators and *tau*, than for $T$. The t-SNE plot for *tau*, Fig. 3a, indicates that if clustering were considered in this single dimension four clusters might be appropriate. However, the property results in Fig. 4 show a clear trend across the two clusters identified by ILS, in clear agreement with that result, when the full feature set is taken into account.

## B. Regression

Each of the two classes were then analysed separately using the ETR, following stratification. Stratification was necessary since the distribution of the property labels in each class is imbalanced. Examples of the stratification for the 25/75 test/train split is shown in Fig. 5 for each class and property label. This process was repeated for each k-fold during our 10-fold cross validation of each model.

The ETR was used to predict the normalised concentration of *Surface Microstructures* and *Surface Facets* for Class 1 and Class 2. In each case, the hyper-parameters were optimised using a grid search, as summarised in Table I.

### 1. *Class 1*

The results of Class 1 reveal similar structure/property relationships for the concentration of *Surface Microstructures* and *Surface Facets*, with some important differences. In Fig. 7 we show the ETR model fit for the training sets (Fig. 7(a,e)), and testing sets (Fig. 7(b,f)), for the *Surface Microstructures* and *Surface Facets*, respectively. For the *Surface Microstructures* (left column) we obtained a training score of $R^2 = 0.998$, a testing score of $R^2 = 0.976$ and a cross-validation score of $R^2 = 0.98 \pm 0.013$. For the *Surface Facets* (right column) we obtained a training score of $R^2 = 0.999$, a testing score of $R^2 = 0.985$ and a cross-validation score of $R^2 = 0.989 \pm 0.006$. In both cases, these results indicate there is no under-fitting (the model is sufficiently sophisticated to capture the complexity in the data) and minimal over-fitting (the model is not fitting to the noise). This is confirmed by the learning curves shown in Figs. 7(c) and 7(g), respectively.

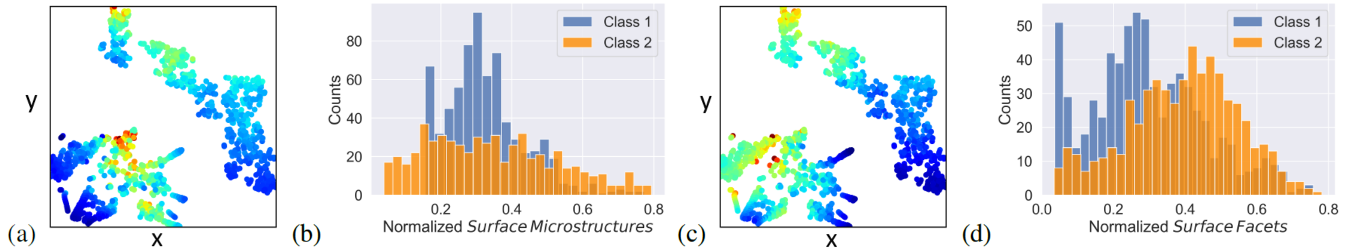

FIG. 4. (a) Distribution of the 1279 platinum nanoparticles (excludes outliers) using t-SNE, colored by the normalized concentration of *Surface Microstructures*, and (b) a histogram of the distribution of *Surface Microstructures*, separated by class. (c) Distribution of the set using t-SNE, colored by the normalized concentration of *Surface Facets*, and (d) a histogram of the distribution of *Surface Facets*, separated by class. Both t-SNE plots are colored from 0 to 1 (from blue to red).

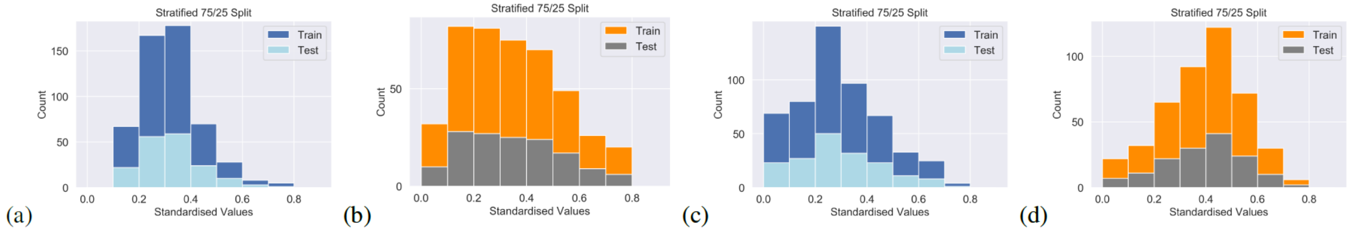

FIG. 5. Stratified 27/75 test/train splits for (a) *Surface Microstructures* in Class 1, (b) *Surface Microstructures* in Class 2, (c) *Surface Facets* in Class 1, and (d) *Surface Facets* in Class 2.

The most useful results come from the feature importance histograms of the *Surface Microstructures* and *Surface Facets*, provided in Figs. 7(d) and 7(h), respectively. These Figs. shows only the top 21 features of the 121 features used to train the models. Here we can see two different structure/property relationships. In the case of the *Surface Microstructures*, the top 5 most important features all relate to coordination numbers (the average coordination number of Pt in the particle, Avg_total; the concentration of Pt atoms with a total coordination of 10, TCN_10; the concentration of Pt atoms with a total coordination of 8, TCN_8; the concentration of Pt atoms with a total coordination of 11, TCN_11; and the concentration of surface atoms with a coordination of 8, SCN_8). These coordination numbers are predominantly indicative of internal disorder (recalling the ideal bulk Pt coordination is 12). The three next most important features all relate to the size of the nanoparticles (the average particle radius, R_avg; the minimum particle radius, R_min; and the total number of bulk-like (non-surface) atoms, N_bulk). An example of a high *Surface Microstructures* Class 1 nanoparticle with atoms coloured by the top 5 coordination numbers is shown in Fig. 6(a).

In the case of the *Surface Facets*, the top 5 most important features all relate to the size of the nanoparticles (the average particle radius, R_avg; the total number of bulk-like (non-surface) atoms, N_bulk; the total number of Pt–Pt bonds, N_bonds; the total number of Pt atoms, N_total; and the total volume of the nanoparticle, Volume). The three next most important features all relate to Pt atoms with a coordination number of 7, (the concentration of surface atoms with a coordination of 7, SCN_7; the concentration of Pt atoms with a total coordination of 7, TCN_7; and the total concentration of Pt atoms a $q6q6$ order parameter of 7, q6q6_T7). These atoms occupy sites on {110} facets surfaces, and the total coordination number (TCN) for a {110} surface atom is the same as the surface coordination number (SCN). An example of a high *Surface Facets* Class 1 nanoparticle with atoms coloured by the top 5 coordination numbers is shown in Fig. 6(b).

### 2. Class 2

The results of Class 2 also show unique structure/property relationships for the concentration of *Surface Microstructures* and *Surface Facets*. In Fig. 7 we show the ETR model fit for the training sets (Fig. 8(a,e)), and testing sets (Fig. 8(b,f)), for the *Surface Microstructures* and *Surface Facets*, respectively. For the Class 2 *Surface Microstructures* we obtained a training score of $R^2 = 0.998$, a testing score of $R^2 = 0.888$ and a cross-validation score of $R^2 = 0.983 \pm 0.013$. For the *Surface Facets* we obtained a training score of $R^2 = 0.999$, a testing score of $R^2 = 0.979$ and a cross-validation score of $R^2 = 0.97 \pm 0.034$. Although the *Surface Microstructures* model gave a lower testing score than the *Surface Facets* model (and the models for Class 1) the learning curves shown in Figs. 8(c) and 8(g) attest to minimal over-fitting.

Turning to the feature importance histograms of the *Surface Microstructures* and *Surface Facets* provided in Figs. 8(d) and 8(h), respectively, we can see that the structure of the surface is much more important in this structurally ordered Class. In the case of the *Surface Microstructures*, half of the

TABLE I. Hyper-parameters optimized using a grid search for each class and property label, using the Extra Trees Regressor.

| Class | Property Indicator | Hyper-parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | criterion | max_depth | min_samples_leaf | min_samples_split | n_estimators | max_features | oob_score |
| 1 | *Surface Microstructures* | mse | None | 2 | 3 | 100 | auto | False |
| 1 | *Surface Facets* | mse | None | 2 | 3 | 200 | auto | False |
| 2 | *Surface Microstructures* | mse | None | 2 | 3 | 500 | auto | False |
| 2 | *Surface Facets* | mse | None | 2 | 3 | 200 | auto | False |

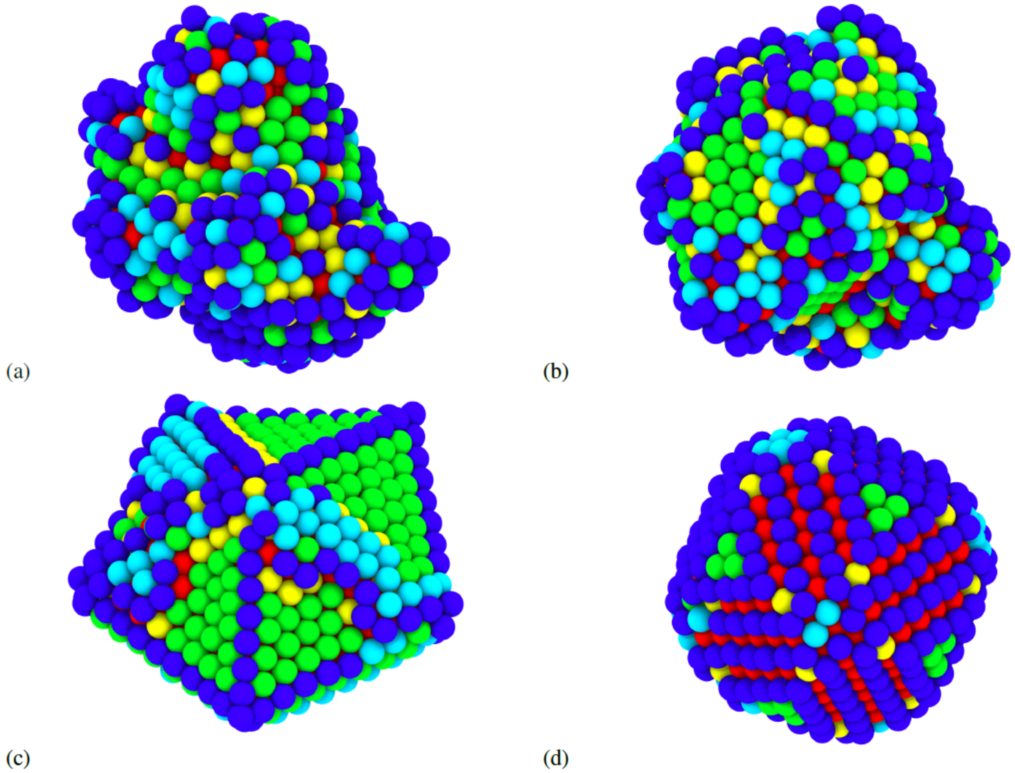

FIG. 6. Examples of Pt nanoparticles in the set, of comparable size, with atoms encoded by the coordination number, for a (a) Class 1 nanoparticle with a high concentration of *Surface Microstructures*, (b) Class 1 nanoparticle with a high concentration of *Surface Facets*, (c) Class 2 nanoparticle with a high concentration of *Surface Microstructures*, and (d) Class 2 nanoparticle with a high concentration of *Surface Facets*. The colouring scheme designates dark blue atoms as having a coordination of 7, light blue 8, green 9, yellow 10 and red 11.

top 8 most important features relate to surface structure (the concentration of surface atoms with a curvature between 1 to 10 degrees, Curve_1-10; the concentration of surface atoms with $q6q6$ order parameter based coordination of 9, q6q6_S9; the concentration surface atoms with a coordination number of 9, SCN_9; and the fraction of atom occupying a {111} facet, S_111). There are three features related to coordination numbers in this top group (Avg_total, SCN_9 and TCN_9) and three order parameters based coordinations (q6q6_T8, q6q6_S9 and q6q6_T9) indicating that it is not just under-coordinated surfaces, but *ordered* under-coordinated surfaces that are important; particularly planar surfaces with a low curvature (Curve_1-10) The number 9 is associated with closed packed surfaces such as the {111} surfaces. An example of a high *Surface Microstructures* Class 2 nanoparticle with atoms coloured by the top 5 coordination numbers is shown in Fig. 6(c).

In the case of the *Surface Facets* for Class 2, half of the top 8 important features also related to the surface structure (SCN_7, q6q6_S7, Curve_31-40 and Curve_21-30), with a strong emphasis on a coordination of 7 (SCN_7, q6q6_T7, q6q6_S7, TCN_7). These features are related to {110} facets, and the following important features contain coordination and order parameter of 11, which are the subsurface atoms along [110] channels. The size is also a consideration (R_min). An example of a high *Surface Facets* Class 2 nanoparticle with atoms coloured by the top 5 coordination numbers is shown in Fig. 6(d), where the {110} surface atoms are shown in dark blue and the sub-{110} surface atoms are shown in red.

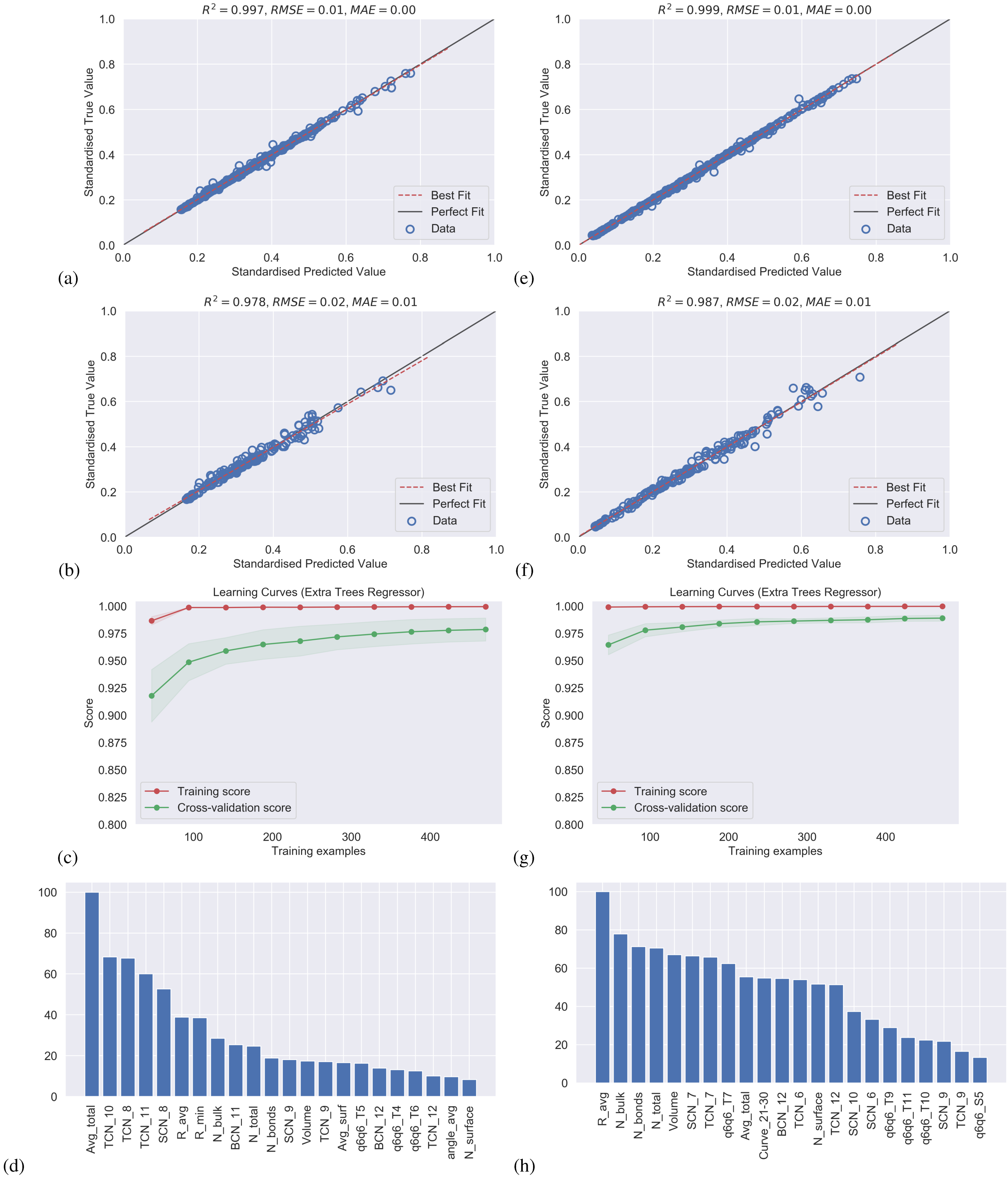

FIG. 7. Results for the extra tree regression models for the *Surface Microstructures* (left) and *Surface Facets* (right), for the disordered Pt nanoparticles of Class 1, including: (a) the *Surface Microstructures* training result, (b) the *Surface Microstructures* testing result, (c) the learning curves for predicting the *Surface Microstructures*, (d) the top 21 structural features determining the concentration of *Surface Microstructures*, (e) the *Surface Facets* training result, (f) the *Surface Facets* testing result, (g) the learning curves for predicting the *Surface Facets*, (h) the top 21 structural features determining the concentration of *Surface Facets*.

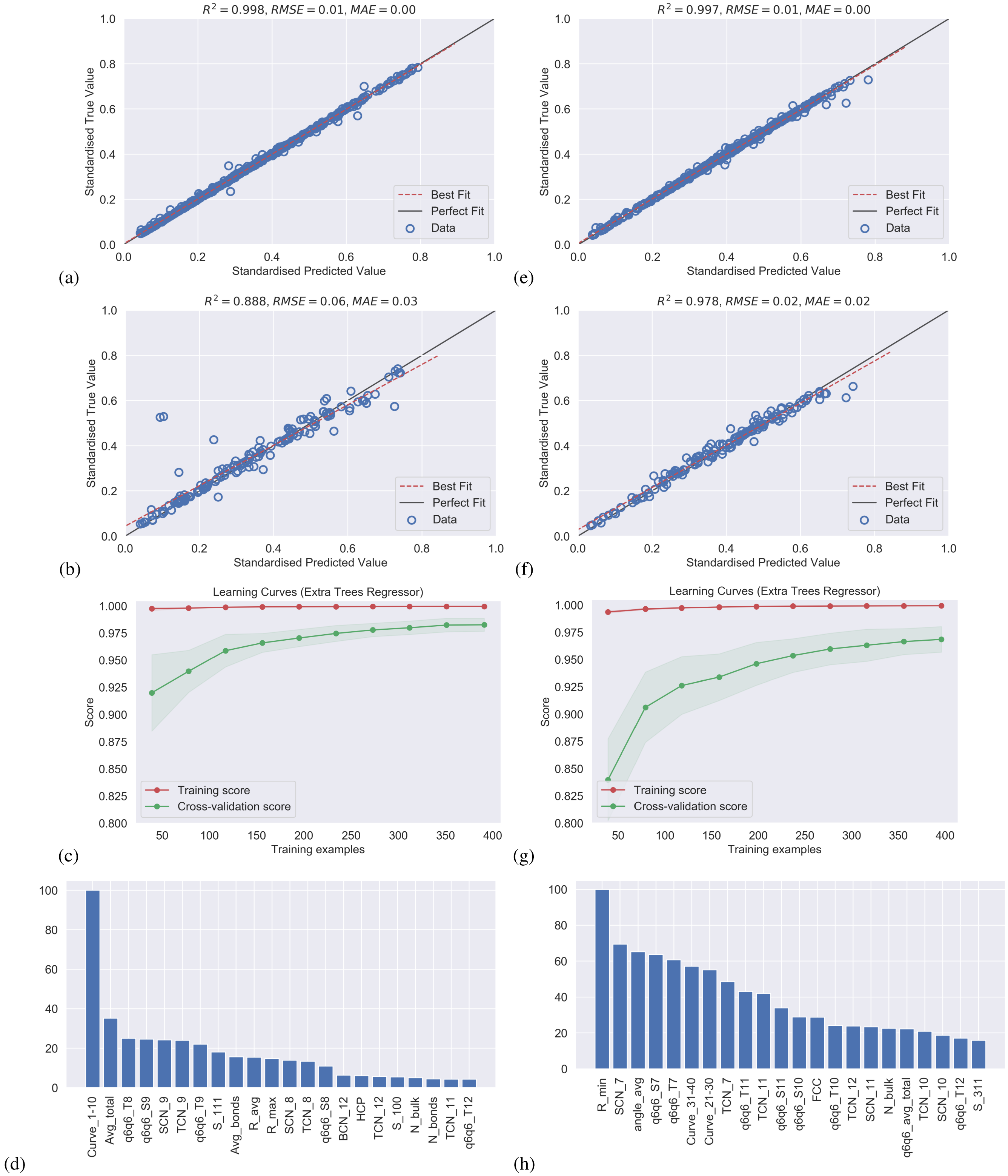

FIG. 8. Results for the extra tree regression models for the *Surface Microstructures* (left) and *Surface Facets* (right), for the ordered Pt nanoparticles of Class 2, including: (a) the *Surface Microstructures* training result, (b) the *Surface Microstructures* testing result, (c) the learning curves for predicting the *Surface Microstructures*, (d) the top 21 structural features determining the concentration of *Surface Microstructures*, (e) the *Surface Facets* training result, (f) the *Surface Facets* testing result, (g) the learning curves for predicting the *Surface Facets*, (h) the top 21 structural features determining the concentration of *Surface Facets*.

## IV. DISCUSSION

Since *Surface Microstructures* are indicative of ORR reactions, and *Surface Facets* are indicative of HER and HOR reactions, these results can be interpreted in terms of their potential impact on catalytic performance.

In Class 1 the results for disordered nanoparticles indicate that ORR efficiency can be controlled by suppressing the crystallinity of the particles (as well as the sizes), which can be done by controlling the *tau* and $T$ (see section on Classification). This is consistent with experimental evidence that amorphous particles display a higher reactivity[71] and complementary statistical analysis.[72] Disordered particles have disordered surfaces, and it has been confirmed using density functional theory[73] that stepped Pt(111) surfaces and nanoparticles with concave features can outperform the activity of flat Pt(111).[74] In this study the authors concluded that concave features can only occur on regular nanoparticles, but the the disordered structures contained in our data set have an enormous fraction of concave atoms. Both steps, and sufficient surface disorder, can increase ORR activity. Other studies have found that (metastable) thin nanorods are also high performing oxygen reduction catalysts,[75] but this sort of structure is not present in our data set where each nanoparticle was optimised as using molecular dynamics during generation. This would be an interesting topic for future work. Our data set is also well beyond the sizes of small clusters that have also been shown to be active.[76]

The results for this class also indicate that HER and HOR efficiency can be controlled by moderating the overall size of disordered nanoparticles (and affecting $\{110\}$ facets), which can also be done by controlling the *tau* and $T$. This is consistent with evidence that disordered amorphous particles have a higher fraction of edge-like atoms, which are known to enhance hydrogen evolution and oxidation reactions[77] and scale as $1/R^2$.

In the case of the ordered Class 2, the *Surface Microstructures* model indicates that ORR efficiency of ordered Pt nanoparticles can be controlled by the moderating of the fraction of flat $\{111\}$ surfaces. This is consistent with experimental observations.[78–80] The *Surface Facets* model indicates that HER and HOR efficiency can be controlled by the moderating of the fraction of $\{110\}$ surfaces, and the overall size. This is also entirely consistent with the experimental observation that HER/HOR is typically an order of magnitude higher on this surface.[79–81]

In both cases, the regressor successfully identifies the right structure/property relationship and highlighted the importance of features that are known to be important experimentally from the entire list of 121. This experimentally consistent and actionable result was only achieved because of the prior classification, that perfectly separated the ordered and disordered particles. If we eliminate this step and apply regression to the entire set, ignoring the classes, the results contain important features for each group combined, as would be expected. Such predictions are confusing and do not provide any logical path as a basis for future work, or a clear way to guide experimental processes. This reinforces the need to apply a strategy of data science and machine learning protocols when seeking to understand complicated structure/property relationships in nanoscience.

## V. CONCLUSIONS

In this study we have used an open data set of ordered and disordered platinum nanoparticles simulated using molecule dynamics to predict the collective structure/property relationship for classes containing distributions of Pt nanoparticles based on their similarity in 121 dimensions. The data set was cleaned and processed to handle redundant features, outliers, normalization and imbalances. Based on clustering using iterative label spreading (ILS), which is well suited to noisy and high-dimensional materials data sets, we identified two clusters that were perfectly separable as classes using the non-linear, non-parametric extra trees classifier. One class contained exclusively disordered nanoparticle, and the other exclusively ordered nanoparticles, which can be separated based on the degree of surface disorder and the growth rate.

Using non-linear, non-parametric extra trees regressors we have subsequently shown that the two classes have different structure/property relationships. Disordered particles (typical of high growth rates and low temperatures) perform better for oxygen reduction reactions if the disorder is increased, and perform better for hydrogen evolution and hydrogen oxidation reactions if the particles are small. Both conditions serve to increase the amount of surface disorder and maximize edge-like atoms. The same machine learning methods identified that ordered nanoparticles will perform better for oxygen reduction reactions if the $\{111\}$ surface area is increased, and will perform better for hydrogen evolution and hydrogen oxidation reactions if the $\{110\}$ surface area is increased. These results agree with experimental observations and support the use of machine learning for multi-structure/multi-property relationships, based on properties averaged over a large sample of structures, rather than specific predictions for individual sizes or shapes that may not be easily controlled in the lab.

## ACKNOWLEDGMENTS

Computational resources for this project have been supplied by CSIRO Scientific Computing.

[1] L. Gou, and C. J. Murphy, Chem. Mater. 2005, **17**, 3668–3672.
[2] B. Wiley, Y. Sun, and Y. Xia, Acc. Chem. Res. 2007, **40**, 1067–1076.
[3] C. Xu, H. Wang, P. K. Shen, and S. P. Jiang, Adv. Mater. 2007, **19**, 4256–4259.
[4] Y. Song, Y. Yang, C. J. Medforth, E. Pereira, A. K. Singh, H. Xu, Y. Jiang, C. J. Brinker, F. V. Swol, and J. A. Shelnutt, J. Am. Chem. Soc. 2004, **126**, 635–645.
[5] S. Tsyganov, J. Keastner, B. Rellinghaus, T. Kauffeldt, F. Westerhoff, and D. Wolf, Phys. Rev. B 2007, **75**, 045421.
[6] G. L. Bezemer,G T. J. Remans, A. P. van Bavel, and A. L. Dugulan, J. Am. Chem. Soc. 2010, **132**, 8540–8541.
[7] A. S. Barnard, H. Konishi, and H. Xu, Catal. Sci. Technol. 2011, **1**, 1440–1488.
[8] A. S. Barnard, and L. Y. Chang, ACS Catal. 2011, **1**, 76–81.
[9] A. S. Barnard, Acc. Chem. Res. 2012, **45**, 1688–1697.

[10] J. Park, J. Joo, S. G. Kwon, Y. Jang, and T. Hyeon, Angew. Chem., Int. Ed. 2007, **46**, 4630–4660.

[11] B. L. V. Prasad, S. I. Stoeva, C. M. Sorensen, and K. J. Klabunde, Chem. Mater. 2003, **15**, 935–942.

[12] S. Stoeva, K. J. Klabunde, C. M. Sorensen, and I. Dragieva, J. Am. Chem. Soc. 2002, **124**, 2305–2311.

[13] D. Lee, R. L. Donkers, J. M. DeSimone, and R. W. Murray, J. Am. Chem. Soc. 2003, **125**, 1182–1183.

[14] Z. Liu, S. Li, Y. Yang, Z. Hu, S. Peng, J. Liang, and Y. Qian, New J. Chem. 2003, **27**, 1748–1752.

[15] N. S. Ramgir, I. S. Mulla, and V. K. Pillai, J. Phys. Chem. B, 2006, **110**, 3995–4001.

[16] Y. Tang, and M. Ouyang, Nature Mater. 2007, **6**, 754–759.

[17] Z. L. Wang, J. Phys. Chem. B 2000, **104**, 1153–1175.

[18] M. Yacaman, K. Heinemann, C. Yang, and H. Poppa, J. Cryst. Growth 1979, **47(2)**, 187–195.

[19] A. Wieckowski, E. R. Savinova, and C. G. Vayenas, Catalysis and Electrocatalysis at Nanoparticle Surfaces, Marcel Dekker, Inc., New York, 2003

[20] M. S. Chen, and D. W. Goodman, Catal. Today, 2006, **111**, 22–33.

[21] E. E. Finney, and R. G. Finke, J. Colloid Interface Sci. 2008, **317**, 351–374.

[22] J. L. Elechiguerra, J. Reyes-Gasga, and M. José-Yacamán, J. Mater. Chem. 2006, **16**, 3906–3919.

[23] S. Maksimuk, X. Teng, and H. Yang, J. Phys. Chem. C, 2007, **111**, 14312–14319.

[24] V. F. Puntes, K. M. Krishnan, and A. P. Alivisatos, Science, 2001, **291**, 2115–2117.

[25] J. Zhang, and J. A. Fang, J. Am. Chem. Soc. 2009, **131**, 18543–18547.

[26] Y.-W. Jun, J.-H. Lee, J.-S. Choi, and J. Cheon, J. Phys. Chem. B, 2005, **109**, 14795–14806.

[27] S. Kumar, and T. Nann, Small, 2006, **2**, 316–329.

[28] T. Teranishi, M. Hosoe, T. Tanaka, and M. Miyake, J. Phys. Chem. B ,1999, **103**, 3818–3827.

[29] T. S. Ahmandi, Z. L. Wang, T. C. Green, A. Henglein, and M. A. El-Sayed, Science, 1996, **272**, 1924–1926.

[30] J. Chen, T. Herricks, and Y. Xia, Angew. Chem., Int. Ed., 2005, **44**, 2589–2592.

[31] H. Lee, S. E. Habas, S. Kweskin, D. Butcher, G. A. Somorjai, and P. Yang, Angew. Chem., Int. Ed. 2006, **45**, 7824–7828.

[32] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nat. Chem. **1**, 37–46 (2009).

[33] V. Stamenkovic, B. S. Mun, K. J. J. Mayrhofer, P. N. Ross, N. M. Marković, J. Rossmeisl, J. Greeley, and J. K. Nørskov, Angew. Chem. Int. Ed. **45**, 2897–2901 (2006).

[34] C. Lu, I. C. Lee, R. I. Masel, A. Wieckowski, and C. Rice, J. Phys. Chem. A, **106**, 3084–3091 (2002).

[35] F. H. B. Lima, J. Zhang, M. H. Shao, K. Sasaki, M. B. Vukmirovic, E. A. Ticianelli, and R. R. Adzic, J. Phys. Chem. C, **111**, 404–410 (2007).

[36] E. Toyoda, R. Jinnouchi, T. Hatanaka, Y. Morimoto, K. Mitsuhara, A. Visikovskiy, and Y. Kido, J. Phys. Chem. C, **115**, 21236–21240 (2011).

[37] Z. Yang, S. Pedireddy, H. K. Lee, Y. Liu, W. W. Tjiu, I. Y. Phang, and X. Y. Ling, Chem. Mater. **28**, 5080–5086 (2016).

[38] A. S. Barnard, B. Motevalli, A. J. Parker, J. M. Fisher, C.A. Feigl, and G. Opletal, Nanoscale **11**, 19190–19201 (2019).

[39] M. Fernandez, A. Bilić, and A.S. Barnard, Nanotech. **28**, 38LT03 (2017).

[40] B. Sun, M. Fernandez, and A. S. Barnard, J. Chem. Info. Mod. **57**, 2413–2423 (2017).

[41] B. Motevalli, B. Sun, and A. S. Barnard, J. Phys. Chem. C, **124**, 7404–7413 (2020).

[42] C. A. Feigl, B. Motevalli, A. J. Parker, B. Sun, and A. S. Barnard, Nanoscale Horiz. **4**, 983–990 (2019).

[43] K. Takahashi, and L. Takahashi, J. Phys. Chem. Lett. **10**, 4063–4068 (2019).

[44] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura, and T. Taniike, ChemCatChem, **11**, 1146–1152 (2018).

[45] P. S. Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abid-Pedersen, and T. Bligaard, ChemCatChem, **11**, 3581–3601 (2019).

[46] H. Barron, and A. S. Barnard, Catal. Sci. Technol. **5**, 2848–2855 (2015).

[47] H. Barron, G. Opletal, R.D. Tilley, and A. S. Barnard, Catal. Sci. Technol. **6**, 144–151 (2016).

[48] V. Mazumder, Y. Lee, and S. Sun, Adv. Funct. Mater. **20**, 1224–1231 (2010).

[49] A. Barnard, and G. Opletal, Platinum Nanoparticle Data Set, v1. CSIRO Data Collection (2019) https://doi.org/10.25919/5d3958d9bf5f7

[50] A. Barnard, B. Sun, and G. Opletal, Disordered Platinum Nanoparticle Data Set, v1. CSIRO Data Collection (2018)

[51] Y. P. Arnaud, Appl. Surf. Sci. **62** 21–35 (1992).

[52] N. Tian, Z. Y. Zhou, and S. G. Sun, J. Phys. Chem. C, **112**, 19801–19817 (2008).

[53] J. S. Spendelow, X. Qinqin, J. D. Goodpaster, P. J. A. Kenis, and A. Wieckowski, J. The Electrochem. Soc. **154**, F238 (2007).

[54] Q. S. Chen, J. Solla-Gullon, S. G. Sun, and J. M. Feliu, Electrochim. Acta, **55**, 7982 (2010).

[55] K. J. J. Mayrhofer, M. Arenz, B. Blizanac, V. Stamenkovic, P. N. Ross, and N. M. Marković, Electrochim. Acta, **50**, 5144–5154 (2005).

[56] N. P. Lebedeva, M. T. M. Koper, J. M. Feliu, and R. A. van Santen, J. Phys. Chem B, **106**, 12938–12947 (2002).

[57] G. Garcia, and M. T. M. Koper, Phys. Chem. Chem. Phys. **10**, 3802 (2008).

[58] G. Garcia, and M. T. Koper, J. Am. Chem. Soc. **131**, 5384 (2009).

[59] Q. S. Chen, F. J. Vidal-Iglesias, J. Solla-Gullon, S. G. Sun, and J. M. Feliu, Chem. Sci. **3**, 136–147 (2012).

[60] H. Barron, G. Opletal, R. D. Tilley, and A. S. Barnard, Nanoscale, **9**, 1502–1510 (2017).

[61] L. M. Falicov, and G. A. Somorjai, Proc. NatI. Acad. Sci. **82** 2207–2211 (1985).

[62] Z. Zhao , Z. Chen, X. Zhang, and G. Lu, J. Phys. Chem. C, **120**, 28125–28130 (2016).

[63] R. A. van Santen, Modern Heterogeneous Catalysis: An Introduction, Wiley-VCH (2017).

[64] M. Fernandez, H. Barron, and A. S. Barnard, RSC Advances, **7**, 48962–48971 (2017).

[65] B. Sun, H. Barron, G. Opletal, and A. S. Barnard, J. Phys. Chem. C, **122**, 28085–28093 (2018).

[66] D. Xu, and Y. Tian, Ann. Data Sci. **2**, 165 (2015).

[67] A. J. Parker, and A. S. Barnard, Adv. Theory Simul. **2**, 1900145 (2019).

[68] B. Motevalli, A. J. Parker, B. Sun, and A. S. Barnard, Nano Futures, **3**, 045001 (2019).

[69] L. J. P. Van Der Maaten, and G. E. Hinton, J. Mach. Learning Res. **9**, 2579–2605 (2008).

[70] A. S. Barnard, and G. Opletal, Nanoscale, **11**, 23165–23172 (2019).

[71] H. Yano, I. Arima, M. Watanabe, A. Iiyama, and H. Uchida, J. Electrochem. Soc. **164**, F966–F972 (2017).

[72] B. Sun, H. Barron, B. Wells, G. Opletal, and A. S. Barnard, Nanoscale, **10**, 20393–20404 (2018).

[73] F. Calle-Vallejo, M. D. Pohl, D. Reinisch, D. Loffreda, P. Sautet, and A. S. Bandarenka, Chem. Sci. **8**, 2283–2289 (2017).

[74] N. Hoshi, and M. Nakamura, Electrochem.**86**, 205–213 (2018).

[75] K. Rossi, G. Giacomo Asaraa, and F. Baletto, Phys. Chem. Chem. Phys. **21** 4888–4898 (2019).

[76] H. Zhai, and A. N. Alexandrova, ACS Catal. **7**, 1905–1911 (2017).

[77] C. Zalitis, A. R. Kucernak, J. Sharman, and E. Wright, J. Mater. Chem. A, **5**, 23328–23338 (2017).

[78] N. M. Marković, H. A. Gasteiger, B .N. Grgur, and P. N. Ross, J. Electro. Chem. **467**, 157–163 (1999).

[79] A.Wieckowski, Interfacial Electrochemistry: Theory, Experiment, and Applications. New York: Marcel Dekker, 1999.

[80] Fuel Cell Science: Theory, Fundamentals, and Biocatalysis. Andrzej Wieckowski, Jens Norskov (Eds) John Wiley & Sons. Inc. 2010

[81] N. M. Markovic, B. N. Grgur, and P. N. Ross, J. Phys. Chem. B **101**, 5405–5413 (1997).