Nanoscale Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal

Cite this: DOI: 10.1039/d5nh00683j

Received 8th October 2025,
Accepted 4th November 2025

DOI: 10.1039/d5nh00683j

rsc.li/nanoscale-horizons

Impact of nanoparticle morphologies on property prediction using explainable AI

Tommy Liu and Amanda S. Barnard *

Every decision made during a machine learning pipeline has an impact on the outcome. Feature selection can reduce overfitting and focus models on the attributes that matter most, and sample selection can reduce bias to ensure models recognise patterns comprehensively. eXplainable AI (XAI) can provide quantitative ways of evaluating the impact of these decisions, and help ensure the right data is used for training models predicting structure property relationships. In this paper we explore the use of residual decomposition with Shapely values to identify which nanoparticle shapes are most influential in predicting charge transfer properties of gold nanoparticles and how they impact the ability to predict the properties of the different morphologies.

New concepts

This paper demonstrates a new application of eXplainable AI (XAI) using a technique known as residual decomposition with Shapley values (RSHAP) to evaluate how nanoparticle morphologies impact predictions of charge transfer properties. Unlike traditional physics-based models that excel at extrapolation, machine learning approaches typically require diverse but representative training sets to enable accurate interpolation. This study applies RSHAP to a dataset of gold nanoparticles to quantitatively reveal which morphologies enhance or degrade the prediction of ionisation potential and electron affinity. This method differentiates nanoparticle contributions as "givers" or "takers," identifying morphologies that significantly improve predictive accuracy *versus* those that negatively influence outcomes. The approach provides a granular view of data valuation by decomposing residual predictions into pairwise interactions among samples. This technique contributes novel insights into nanoscience by clarifying the role individual shapes play in predictive models, aiding strategic selection of morphologies for training sets. It specifically underscores that including certain unconventional or polycrystalline shapes may not necessarily degrade predictive performance, challenging typical data exclusion practices and offering an evidence-based approach for optimizing experimental and computational resources in nanotechnology research.

The most important decision when using machine learning to predict the structure–property relationships of nanomaterials is which structures to include in the training set.[1] Unlike conventional physics-based and phenomenological models, which can be capable of reliable extrapolation,[2,3] machine learning excels at interpolation,[4,5] so it is essential that the training set contains a diverse set of samples that comprehensively cover the configuration space.[6–8] Training sets do not need to be exhaustive,[9] but some configuration spaces are combinatorial,[10,11] making it difficult to decide what to exclude without compromising predictive performance. The value of a machine learning-based structure–property relationship is that it can predict the properties of unseen nanomaterials, and this is redundant if almost all samples must be seen to ensure the results are reliable. One way to overcome this dilemma is to undertake an exploratory (preliminary) study using limited numbers of samples with extreme diversity, and measure exactly what each structure is doing to the model and its ability to predict the properties.

eXplainable AI (XAI)[12] provide a suite of post-hoc model-agnostic methods capable of forensic examination of machine learning models.[13] XAI can help researchers understand which structural features are most important to the underlying

ANU School of Computing, 145 Science Road, Acton, Australia.
E-mail: amanda.s.barnard@anu.edu.au

Amanda S. Barnard

After a decade of success Nano Horizons as grown into an important venue for cutting edge research in nanoscience and nanotechnology. As Nano Horizons enters the next decade, on the wave of artificial intelligence, it is our pleasure to contribute this new research demonstrating how the choice of nanomaterials made in the computer to the lab underpins our ability to predict new materials in the future.

prediction,[14] regardless of the model architecture, and therefore how removing structural features during data pre-processing impacts the outcome.[15–17] XAI can also help researchers understand which individual structures are most influential,[18,19] and how decisions to remove outliers or restrict the configurations space to sub-set of samples affects property predictions. This can assist in data valuation, and inform data acquisition to insure that costly or time-consuming experiment are focused on structures that improve performance. Recently a new method known as RSHAP was reported[20,21] that decomposes the residual of model predictions to explain how sample instances contribute to the prediction of themselves and others, and how choosing the right data can make a difference.

A long standing topic in nanoscience has been the relationship between the morphology of nanoparticles and their properties.[22–28] It has been well-established that some properties are shape-dependent,[29–32] and there is compelling evidence that other properties are affected by the overall shape,[33,34] particularly those related to the surfaces.[35,36] In these cases it is clear that a diverse range of shapes should be included in predictive studies to capture latent relationships, but it is unclear which shapes and how many of them. In this study we apply the RSHAP approach to a modest set of gold nanoparticles to identify which morphologies contribute most to the residual of models predicting the ionisation potential (IP) and the electron affinity (EA), using a public data set originally generated with electronic structure simulations. We compare models trained using features describing the structure of entire nanoparticles or those describing just the surfaces, and find that different shapes can improve overall model performance and the ability to accurately predict the charge transfer properties of other shapes.

The nanoparticle data set[37] used here contains 2248 gold nanoparticles, but was not generated as part of this study. In this study we use the 691 sample structures that have been labelled with IP and EA ranging in size from 13 atoms to 2479 atoms in size, described by a range of manually extracted features outlined in the metadata. The feature space includes the number of Au atoms with coordination numbers (CN), generalised coordination number (GCN) and $q6q6$ order parameters, calculated using the NCPac software[38] based on the total atoms in the nanoparticle (T), the bulk atoms (B) and the surface atoms (S). The method for calculating these features is reported elsewhere.[39] The feature space also contains various bond lengths and angles which would be common to all groups and therefore not used in this study. For the purposes of this demonstration, we have used subsets of features for T and S feature groups (descriptors), which are entirely disjointed. The B descriptor has been omitted as it is assumed interior atoms have little or no impact on surface charge transfer properties, and a comparison of T and S will be sufficient to determine if models predicting surface properties should be trained exclusively with surface features. This assumption will only hold for samples that have a significant number of interior atoms. For example, it has been reported that in the case of Nb, structural isomerism has a very strong effect on the reactivity of small Nb_9 and Nb_{12} clusters[40] (which have no interior atoms), which would invalidate this assumption. In the present study only one sample (out of 691) was characterised in the *meta* data as "all surface" (bulk atom coordination number is NaN). Details of the data set, the feature space and descriptors are provided in the SI.

Table 1 External labels used for annotation (not used for training). N = the population in the data set following data cleaning (in %). m = modified (chamfered edge), t = minimally truncated vertex (removal of vertex atoms). * Contains twinning and/or stacking faults (358 nanoparticles), representing 52% of the data set

ID	Morphology	Facets	N
C	Cube (hexahedron)	{100}	11
CO	Cuboctahedron	{100}, {111}	11
DH*	Decahedron	{110}, {111}	152
GRC	Great rhombicuboctahedron	{100}, {110}, {111}	13
HO	Hexoctahedron	{123}	24
IH*	Icosahedron	{111}	5
OH	Octahedron	{111}	10
POLY*	Irregular polycrystalline particle	Various	206
RD	Rhombic dodecahedron	{110}	9
RH	Rhombi-truncated hexahedron	{100}, {110}	22
RO	Rhombi-truncated octahedron	{110}, {111}	9
SRC	Small rhombicuboctahedron	{100}, {110}, {111}	15
T	Tetrahedron	{111}	11
TC	Truncated cube	{111}	8
TH	Tetrahexahedron	{210}	10
TO	Truncated octahedron	{100}	19
TR	Trisoctahedron	{331}	15
TZ	Trapezohedron	{311}	9
mTO	Modified truncated octahedron	{100}, {110}, {111}	26
tHO	Truncated hexoctahedron	{123}	29
tRD	Truncated rhombic dodecahedron	{110}	7
tT	Truncated tetrahedron	{111}	21
tTH	Truncated tetrahexahedron	{210}	22
tTR	Truncated trisoctahedron	{331}	5
tTZ	Truncated trapezohedron	{311}	22

The morphology identifiers (IDs) are detailed in Table 1, along with the population of each of these shapes. Each nanoparticle in the data set is annotated by a morphology identifier as an external label that is not used for training, and the distribution of the charge transfer properties for each morphology is shown in Fig. 1.

We have trained regression models to predict the IP and EA for the T and S descriptors. Linear Ridge regression[41] was compared to XGBoost[42] and found to be superior for each descriptor and target label. Details of the methods, model tuning and hyperparameters are provided in the SI, but the final model scores are listed in Table 2. Fig. 2(a) and (b) show the IP and EA parity plots for the testing set with the T descriptor, and Fig. 3(a) and (b) show the IP and EA parity plots for the S descriptor, respectively. In each case the points are annotated by the nanoparticle morphology. The learning curves, and feature importance profiles are provided in the SI. The testing results (Table 2) show that the model scores are imperfect, with numerous samples having significant residuals (see SI). The charge transfer properties for some morphologies are more difficult to predict than others, regardless of their values or distributions.

A better understanding of the influence of certain samples and morphologies can be achieved using concepts from

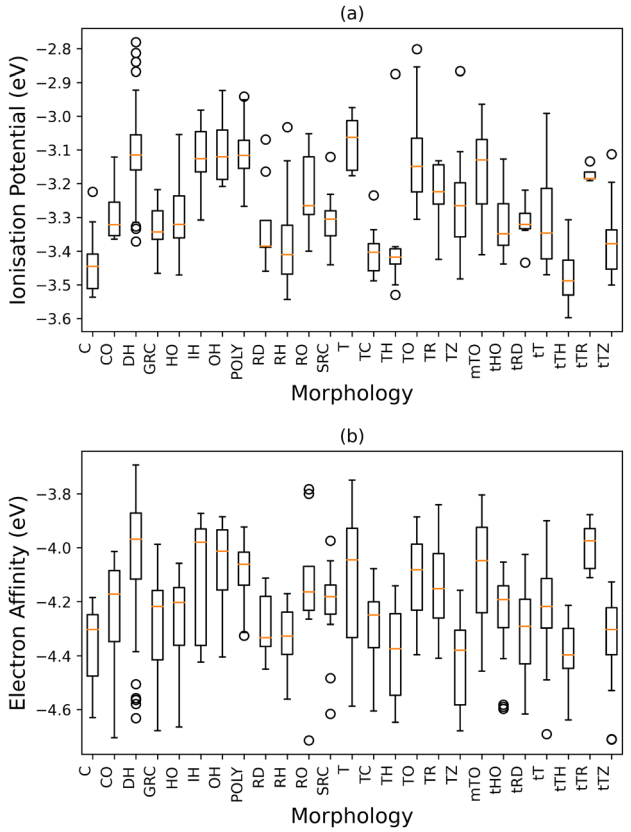

Fig. 1 Distribution of (a) the ionisation potential (IP) and (b) the electron affinity (EA) for each of the Au nanoparticle morphologies listed in Table 1, following data cleaning to remove outliers.

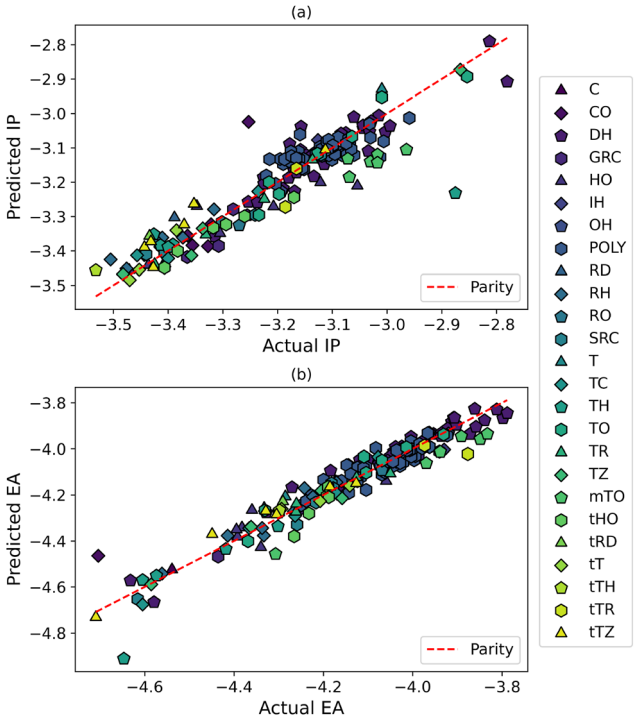

Fig. 2 Model testing results predicting (a) the ionisation potential (IP) and (b) the electron affinity (EA) for Au nanoparticle morphologies listed in Table 1, using the T descriptor describing the structure of the entire nanoparticles.

Table 2 Ridge regressor model performance trained with each descriptor, predicting each charge transfer property

	Ionisation potential (IP)			Electron affinity (EA)		
Descriptor	MAE	RMSE	R^2	MAE	RMSE	R^2
T (all)	0.041	0.003	0.863	0.039	0.003	0.922
S (surface)	0.048	0.004	0.818	0.048	0.004	0.890

cooperative game theory such as computing the Shapley values (ϕ_i),[43,44] as described by:

$$\phi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{i\}) - v(S)]. \quad (1)$$

where F is the set of all samples, i is an individual nanoparticle $i \in F$, and $v(\cdot)$ is the cooperative game (the loss). Shapley values are the weighted overage over all possible subsets that do not contain i and the marginal effect of i is measured by $v(S \cup \{i\}) - v(S)$. One of the core concepts around using Shapley values is the additive property, where the sum of the Shapley values of each individual sample i sums to the value of the set, that is $\sum_{i=1}^{n} \phi_i = v(F)$. By solving for ϕ_i for each instance, we can identify the nanoparticles most responsible for improving model accuracy.[45,46] This can inform which types of new data instances we should generate or sample to produce better models; which nanoparticle shapes are more "valuable."

The residual decomposition framework for Shapley values (RSHAP) extends the concept of data value[20] to consider the pairwise effect of each sample instance on other instances (in the context of the model), in terms of their contribution and composition (CC). The contribution measures how much an individual sample affects the predicted outcomes of other samples, and the composition measures how other samples affected the model prediction for a given sample. These "CC" effects are calculated by setting the value function $v(\cdot)$ to be the impact that a sample x_i has on the predicted outcomes of all other samples in $F \setminus \{x_i\}$, and is precisely evaluated using the residual values over the entire set as:

$$v(S) = \{f_S(x_i) - y_i\}_{i=1}^{n}. \quad (2)$$

The resultant contribution-composition matrix ("CC-matrix," Φ) contains rows of Shapley values ϕ_i for each ith nanoparticle, and n values predicting how much i affects the prediction of all n nanoparticles (including i). A simple interpretation of a CC-plot is outlined in Fig. 4, where positive Contribution values indicate that instances work to make the model worse, and negative contributions work to improve the performance of the model. This method has recently been used to explore the impact of specific chemical elements on the prediction of properties of dilute solutes, perovskites, and metallic glasses,[21,47] presenting the CC-matrix *via* "CC-plots," and the

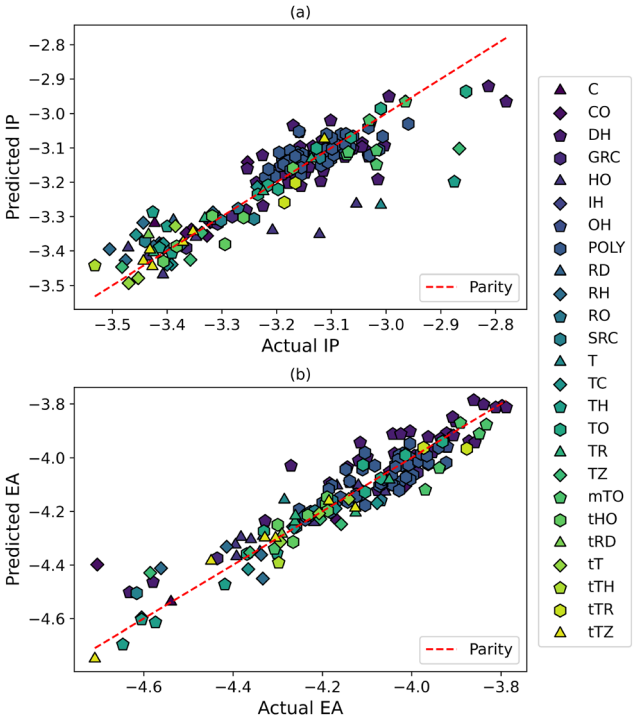

Fig. 3 Model testing results predicting (a) the ionisation potential (IP) and (b) the electron affinity (EA) for Au nanoparticle morphologies listed in Table 1, using the *S* descriptor describing the structure of the surfaces of the nanoparticles.

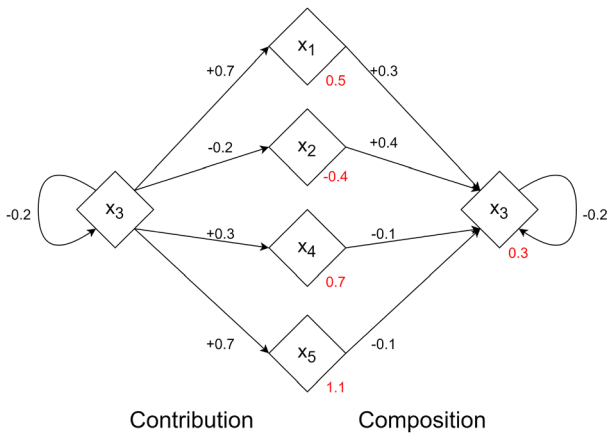

Fig. 4 An example of contribution and composition values for an example dataset of five samples. Red values indicate the residual value the model produces for that instance, black values indicate the contribution effect that an instance has upon another. Figure reproduced from Liu and Barnard under a CC 4.0 BY Deed license.

pairwise interactions *via* heatmaps. Code and notebooks to reproduce these results are provided at ref. 48.

The CC-plots for the *T* descriptor generated using all Au atoms in each particle are shown in Fig. 5 and 6, for the IP and EA of the testing sets, respectively. Firstly, we see that the majority of the samples lie around the origin indicating that their overall impact is relatively low. In Fig. 5(a) and 6(a) the CC-plots are annotated by the charge transfer properties, and we can see that higher energies occupy the upper left or the lower right quadrant. The upper left quadrant contains samples that contribute more to the residuals, making the prediction of the IP or EA worse, but with low residuals of their own. This can occur when data that does not fit the trend of the model or arose from a different distribution (outliers) and we can think of them as the "takers." The lower right quadrant contains samples that have a negative contribution to the residuals, improving the models, but having a higher residual themselves. We can think of them as the "givers." In contrast, a large number of low IP and EA samples reside in the upper right quadrant, where the particles have a high residual, and increase the model residuals on other samples. These particles significantly reduce model performance. Finally, in the lower left quadrant are the samples that have low residuals and decrease the residuals on other samples. These nanoparticles significantly improve model performance.

In Fig. 5(b) and 6(b) the CC-plots are annotated by the morphologies, where we can see that the undesirable morphologies is the upper right quadrant include the highly faceted tTZ, tTR and tTH. These are shapes have 24 high index facets with truncated vertices,[49,50] resulting in 38 facets. These facets are high energy planes and these shapes are usually omitted from most studies due to the reduced thermodynamic stability; a decision that is supported by the residual decomposition that indicates they decrease model performance even when the formation energetics are not considered. Very few of these shapes occupy the desirable lower left quadrant of morphologies with low residual that also reduce the residuals on other nanoparticles. When using the *T* descriptors calculated using all Au atoms most desirable shape to include to improve the prediction of the IP is the octahedron (OH) and the most desirable shape to include to improve the prediction of the EA is the cube (C); both shapes that are commonly included in computational nanoscience research.[51]

The CC-plots for the *S* descriptor generated using the surface Au atoms in each particle are shown in Fig. 7 and 8, for the IP and EA of the testing sets, respectively. Although the distribution of the overall CC-plots for the *S* descriptor group are similar to the *T* group, the separation of the high IP and EA nanoparticles into givers and takers is more distinct, though there are far more givers (of higher residuals) than takers. There are also far more nanoparticles in the undesirable upper right quadrant, and far fewer in the desirable lower left quadrant. Comparing the morphology-annotated Fig. 7(a) and 8(b) with Fig. 5(b) and 6(b) we can see highly faceted near-spherical nanoparticle are more centralised in the CC-plots, indicating that the impact of these shapes on the model residuals is mitigated by using descriptors based only on the surface atoms.

Regardless of the descriptor, POLY samples, which include a wide variety of shapes with numerous internal point defects, twins, staking faults, surface facets, terraces, edges, kinks, vertices and protrusions,[52] are rarely outliers in the CC-plots. The "teal hexagon" annotation are most tightly packed in the centre of the CC-plots, indicating that the impact of POLY on

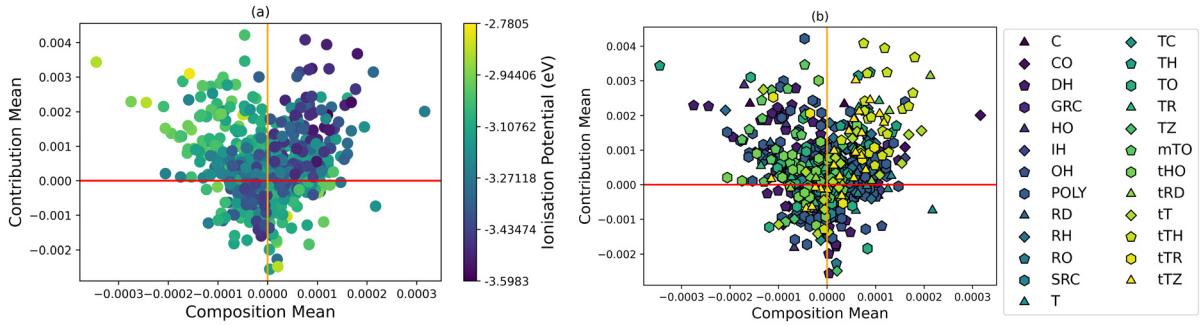

Fig. 5 CC-plots for the Ridge regressor trained using the *T* descriptor predicting the ionisation potential (IP), annotated by (a) the IP in eV, and (b) the particle morphology.

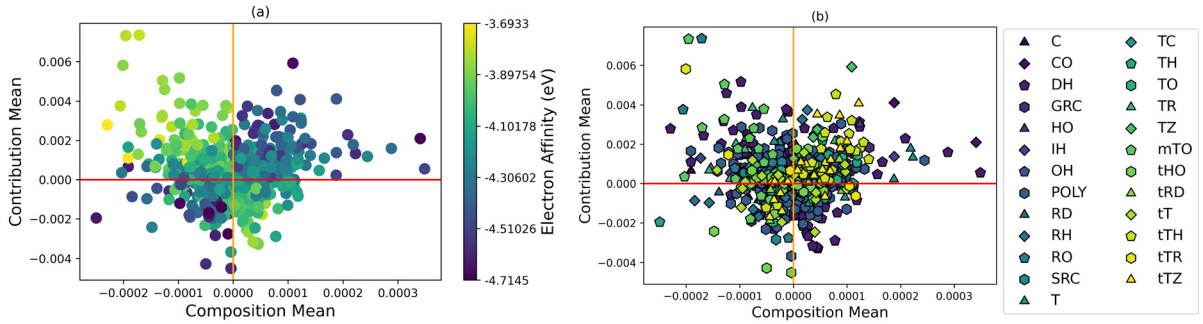

Fig. 6 CC-plots for the Ridge regressor trained using the *T* descriptor predicting the electron affinity (EA), annotated by (a) the EA in eV, and (b) the particle morphology.

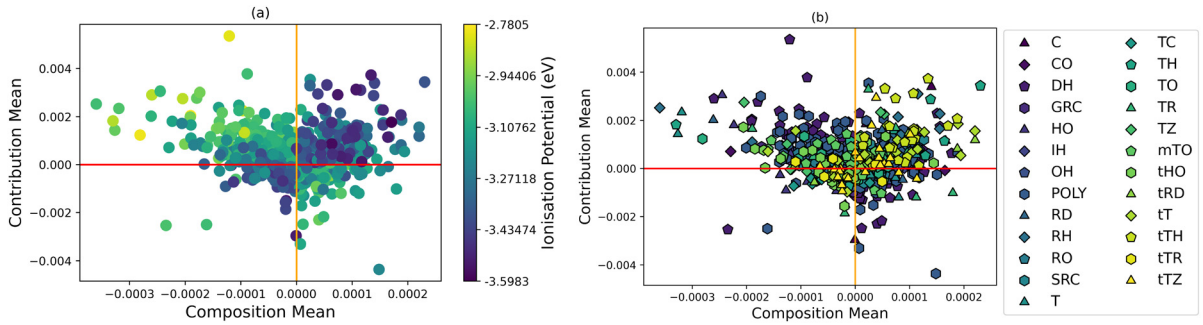

Fig. 7 CC-plots for the Ridge regressor trained using the *S* descriptor predicting the ionisation potential (IP), annotated by (a) the IP in eV, and (b) the particle morphology.

the models, and the predictions of more perfect zonohedrons, is low. Polycrystalline particles are often excluded from studies of nanomorphology and machine learning of nanoparticles, usually due to reduced thermodynamic stability[53] with respect to regular zonohedrons, but the present results suggest these decisions may be unfounded. In this case including a diverse mix of irregular morphologies does not significantly degrade model performance.

The CC values can also be used to analyse the pairwise effects of samples in the form of a heatmap. A CC-heatmap shows how much each morphology contributes to the residuals across the rows and the composition of each of the residuals in the columns. Individual cells in the heatmap based on some morphology (scaled to $[-1, 1]$) represent how much the particular shape and structure of nanoparticle contributes to our ability to accurately predict the charge transfer properties of others. This is achieved by changing the Shapley valuation function and normalisation to measure contribution values using the residual values e_i over each of the samples, as given by:

$$v(S) = L(f_S(X), Y). \quad (3)$$

Fig. 9(a) and (b) show the pairwise CC-heatmaps for models trained using the *T* descriptor to predict the IP and EA of the testing sets, respectively. The order of the morphologies down the rows and across the column has been changed to reflect the average contribution to the (respective) model residuals;

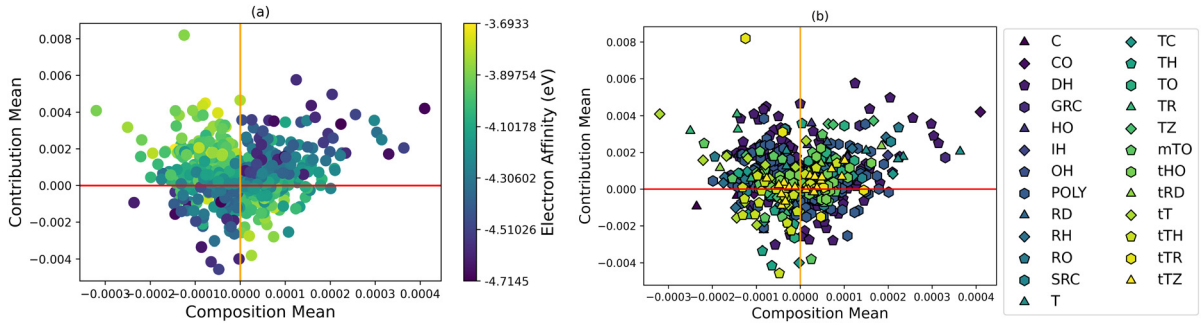

Fig. 8 CC-plots for the Ridge regressor trained using the *S* descriptor predicting the electron affinity (EA), annotated by (a) the EA in eV, and (b) the particle morphology.

morphologies at the top (and left) lower the residuals on other morphologies and those at the bottom (and right) increase the model residuals. There are a few morphologies that stand out. In the IP model (Fig. 9(a)) the tetrahedron (T) lowers the residual on most shapes, but increases the residual on the rhombitruncated octahedron (RO). The icosahedron (IH) significantly increased the residual on the hexoctahedron (HO), but there are very few of each of these shapes in the data set (see Table 1). The tetrahexahedron (TH) increases the residual on the cuboctahedron (CO) and the cube (C), and the truncated tetrahexaherdon (tTH) increases the residual on numerous shape including the CO and C. In the EA model (Fig. 9(b)) many more morphologies increase the residuals on others. In particular the IH increases the residual on the truncated tetrahedron (tT), the T increases the residual on the five-fold twinned decahedron (DH) and the truncated trisoctahedron (tTR) increases the residual on the trapezohedron. The most desirable shape, the C, only increases the residual on itself.

Fig. 10(a) and (b) show the pairwise CC-heatmaps for models trained using the *S* descriptor to predict the IP and EA of the testing sets, respectively, ordered in the same way as Fig. 9. In the case of the IP model (Fig. 10(a)) the T significantly increases the residuals, but only on itself, which is likely due to the impact of the highly acute edges and vertices that are unique to this shape and enhanced when using the *S* descriptor. There is no harm to keeping this shape even though the residuals are high. The predictions of the modified truncated octahedron (mTO) and the rhombi-truncated hexahedron (RH) are degraded by the tTH and the TH. In the case of the EA model (Fig. 10(b)), the T, trapezohedron (TZ), tTR and truncated rhombidodecahedron (tRD) increased the residuals on the RH, small rhombicuboctahedron (SRC), tTR and TR, respectively.

Overall these results have shown that the choice of which nanoparticle morphologies to include in data sets for machine learning can impact the outcome, and that impact is not evenly distributed. Certain shapes have higher residuals, degrading model performance, and should be avoided. Some shapes increase the residuals on others, regardless of their own residuals, and should also be avoided. Other shapes have low residuals, and can even lower the residuals on others, making them very useful and ideal candidates to increase the size of a data set. Depending on the focus of a given study, shapes can be combined strategically to improve overall predictive ability or to mitigate individual effects. For example, is the aim is to study the EA of shapes enclosed entirely by {111} facets using the entire nanoparticle (T descriptor) Fig. 9(b) indicates that the IH increases the residual on tT, but this can be mitigated by adding more T. The addition of T will increase the residual on

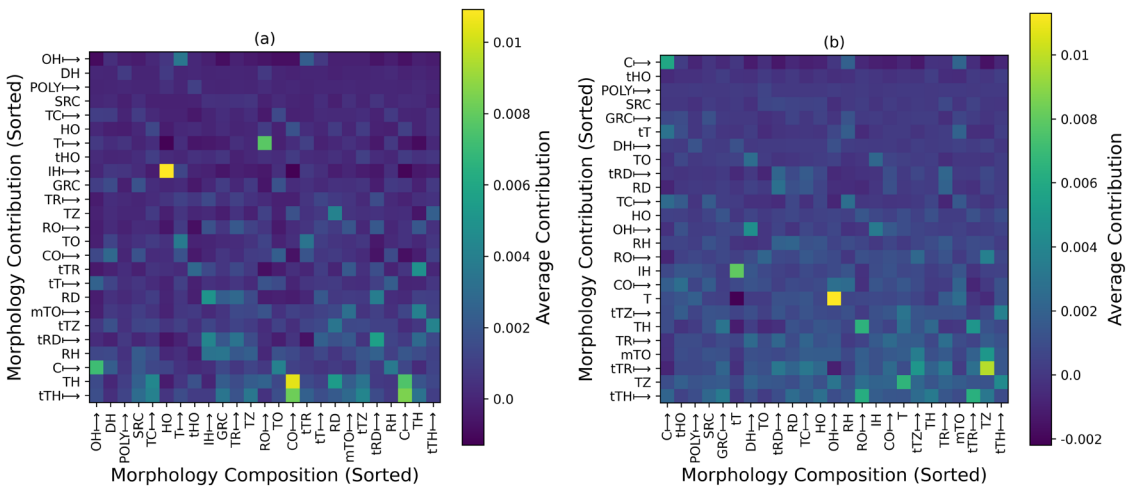

Fig. 9 CC-heatmaps for the Ridge regressor trained using the *T* descriptor predicting (a) the ionisation potential (IP), and (b) the electron affinity (EA).

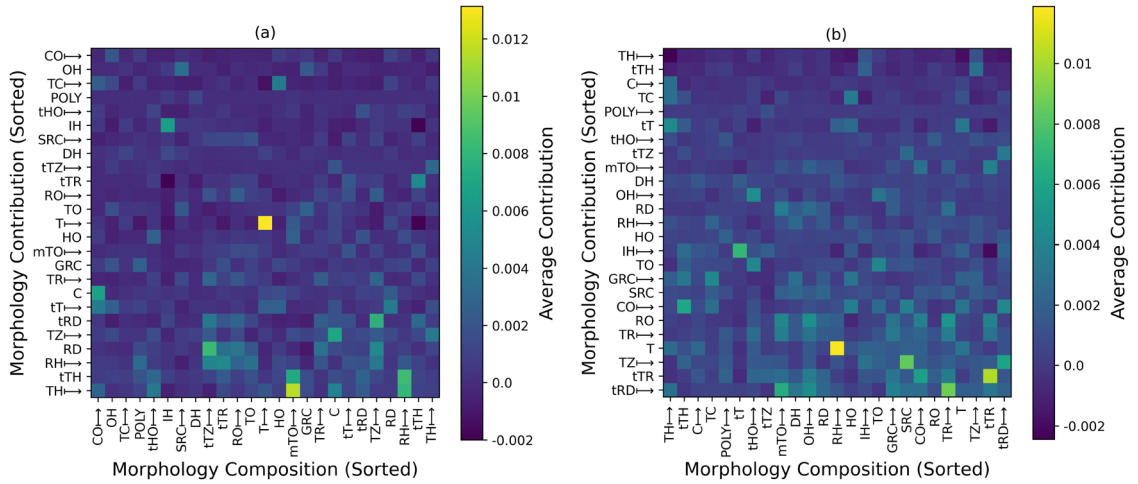

Fig. 10 CC-heatmaps for the Ridge regressor trained using the *S* descriptor predicting (a) the ionisation potential (IP), and (b) the electron affinity (EA).

OH, but this can be mitigated including less OH or adding one shape with {111} facets, such as the tTR (which has a shape very similar to the OH).

Given some of these relationships will not be known in advance, and they depend on the features used in the training set and the target property label, the best approach is to start with a minimal set of representative morphologies, train a model and then explain it with RSHAP. Depending on the outcome strategic decisions can be made as to which shapes to add to achieve the goals of the study. The data set can easily be iteratively expanded, or incorporated into an active learning pipeline.[54] This approach is general and can be applied to other tasks, models and nanomaterial systems. Understanding how the choice of data impacts final model and and its abilities leads to better decision making in the beginning, and better explanations at the end.

Author contributions

TL: conceptualisation, methodology, writing – original draft, investigation, software, formal analysis, validation, visualisation. ASB: conceptualisation, writing – review & editing, supervision, project administration, data curation, resources.

Conflicts of interest

There are no conflicts to declare.

Data availability

Additional information supporting this article has been uploaded as part of the supplementary information (SI). Supplementary information: detailed list of the descriptors, final model hyperparameters from model optimisation, learning curves from model training, residuals from model testing. See DOI: **https://doi.org/10.1039/d5nh00683j**.

The data set is available at: **https://doi.org/10.25919/bzag-6w95**. The software is available at: **https://doi.org/10.6084/m9.figshare.30254461.v6**.

Acknowledgements

This research was supported by the National Computational Infrastructure (NCI) under project p00, and the Australian Government Research Training Program (RTP) Scholarship.

References

1 A. S. Barnard, B. Motevalli, A. J. Parker, J. M. Fischer, C. A. Feigl and G. Opletal, *Nanoscale*, 2019, **11**, 19190–19201.
2 A. Pastore and M. Carnini, *J. Phys. G*, 2021, **48**, 084001.
3 A. Rajput, G. Shevalkar, K. Pardeshi and P. Pingale, *OpenNano*, 2023, **12**, 100147.
4 I. H. Sarker, *SN Comput. Sci.*, 2021, **2**, 1–21.
5 C. Malica, K. Novoselov, A. S. Barnard, S. V. Kalinin, S. R. Spurgeon, K. Reuter, M. Alducin, V. L. Deringer, G. Csanyi, N. Marzari, S. Huang, G. Cuniberti, Q. Deng, P. Ordejon, I. Cole, K. Choudhary, K. Hippalgaonkar, R. Zhu, O. A. von Lilienfeld, M. Hibat-Allah, J. Carrasquilla Alvarez, G. Cisotto, A. Zancanaro, W. Wenzel, A. C. Ferrari, A. Ustyuzhanin and S. Roche, *J. Phys.*, 2025, **8**, 021001.
6 G. Fenza, M. Gallo, V. Loia, F. J. Orciuoli and E. E. Herrera-Viedma, *Appl. Soft Comput.*, 2021, **106**, 107366.
7 Y. Gong, G. Liu, Y. Xue, R. Li and L. Meng, *Inf. Softw. Technol.*, 2023, **162**, 107268.
8 D. Schwabe, K. Becker, M. Seyferth, A. Klass and T. Schäffter, *NPJ Digital Med.*, 2024, **7**, 2–3.
9 K. Li, D. Persaud, K. Choudhary, B. L. DeCost, M. Greenwood and J. R. Hattrick-Simpers, *Nat. Commun.*, 2023, **14**, 7283.
10 R. C. Pullar, in Combinatorial Materials Science, and a Perspective on Challenges in Data Acquisition, *Analysis and Presentation*, ed. T. Lookman, F. J. Alexander and

K. Rajan, Springer International Publishing, Cham, 2016, pp. 241–270.
11 A. Ludwig, *npj Comput. Mater.*, 2019, **5**, 1–7.
12 T. Liu and A. S. Barnard, *Cell Rep. Phys. Sci.*, 2023, **4**, 101630.
13 S. Li, X. Wang and A. S. Barnard, *Mach. Learn.*, 2024, **6**, 013002.
14 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
15 A. Mangal and E. A. Holm, *Integ. Mater. Manuf. Innovat.*, 2018, **7**, 87–95.
16 C. A. Rickert, M. Henkel and O. Lieleg, *APL Mach. Learn.*, 2023, **1**, 016105.
17 Y. Hu, R. Sandt and R. Spatschek, *Sci. Rep.*, 2024, **14**, 20449.
18 A. S. Barnard, *Cell Rep. Phys. Sci.*, 2021, **3**, 100696.
19 A. S. Barnard and B. L. Fox, *Chem. Mater.*, 2023, **35**, 8840–8856.
20 T. Liu and A. S. Barnard, International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, 2023, pp. 21375–21387.
21 T. Liu, Z. Y. Tho and A. S. Barnard, *Digital Discovery*, 2024, **3**, 422–435.
22 G. C. Schatz, K. L. Kelly, E. A. Coronado, L. L. Zhao, D. Beljonne, C. Curutchet, G. D. Scholes, B. O. Dabbousi, J. Rodríguez-Viejo, F. V. Mikulec, C. J. Murphy, T. K. Sau, A. Gole, C. J. Orendorff, J. Gao, L. Gou and S. E. Hunyadi, *J. Phys. Chem. B*, 2003, **107**, 668–677.
23 A. S. Barnard and I. K. Snook, *J. Chem. Phys.*, 2004, **120**, 3817–3821.
24 A. S. Barnard and L. A. Curtiss, *Rev. Adv. Mater. Sci.*, 2005, **10**, 105–109.
25 A. S. Barnard, *J. Mater. Chem.*, 2006, **16**, 813–815.
26 T. Gomathi, K. Rajeshwari, V. Kanchana, P. N. Sudha and K. Parthasarathy, in *Impact of Nanoparticle Shape, Size, and Properties of the Sustainable Nanocomposites*, ed. Inamuddin, S. Thomas, R. Kumar Mishra and A. M. Asiri, Springer International Publishing, Cham, 2019, pp. 313–336.
27 M. C. Arno, M. Inam, A. C. Weems, Z. Li, A. L. A. Binch, C. I. Platt, S. M. Richardson, J. A. Hoyland, A. P. Dove and R. K. O'Reilly, *Nat. Commun.*, 2020, **11**, 1420.
28 R. Ridolfo, S. Tavakoli, V. Junnuthula, D. S. Williams, A. Urtti and J. C. M. van Hest, *Biomacromolecules*, 2020, **22**, 126–133.
29 S. Mostafa, F. Behafarid, J. R. Croy, L. K. Ono, L. Li, J. C. Yang, A. I. Frenkel and B. R. Cuenya, *J. Am. Chem. Soc.*, 2010, **132**, 15714–15719.
30 R. Essajai, Y. Benhouria, A. Rachadi, M. Qjani, A. Mzerd and N. Hassanain, *RSC Adv.*, 2019, **9**, 22057–22063.
31 H. Shi, A. S. Barnard and I. K. Snook, *Nanoscale*, 2012, **4**, 6761–6767.
32 J. Auclair and F. Gagné, *Nanomaterials*, 2022, **12**, 3107.
33 D. V. Kladko, A. S. Falchevskaya, N. Serov and A. Y. Prilepskii, *Int. J. Mol. Sci.*, 2021, **22**, 5266.
34 K. Öztürk, M. Y. Kaplan and S. Calis, *Int. J. Pharm.*, 2024, **666**, 124799.
35 H. Lee, *RSC Adv.*, 2014, **4**, 41017–41027.
36 K. An and G. A. Somorjai, *ChemCatChem*, 2012, **4**, 1512–1524.
37 G. Opletal and A. S. Barnard, Au Nanoparticle Data Set, CSIRO Data Collection, 2024, **v2**, DOI: **10.25919/bzag-6w95**.
38 G. Opletal, J. Ting and A. S. Barnard, *NCPac*, CSIRO Sofware, 2024, **v1**, DOI: **10.25919/tfv3-he58**.
39 J. Y. C. Ting, G. Opletal and A. S. Barnard, *Catal. Sci. Technol.*, 2024, **14**, 6651–6661.
40 M. B. Knickelbein and S. Yang, *J. Chem. Phys.*, 1990, **93**, 5760–5767.
41 D. N. Schreiber-Gregory, *Model. Assist. Stat. Appl.*, 2018, **13**, 359–365.
42 T. Chen and C. Guestrin, *A Scalable Tree Boosting System*, New York, NY, USA, 2016, pp. 785–794.
43 S. M. Lundberg and S. Lee, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017, pp. 4765–4774.
44 C. Molnar, *Interpretable Machine Learning*, 2nd edn, 2022.
45 A. Ghorbani and J. Y. Zou, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, 2019, pp. 2242–2251.
46 S. Tang, A. Ghorbani, R. Yamashita, S. Rehman, J. A. Dunnmon, J. Zou and D. L. Rubin, *Sci. Rep.*, 2021, **11**, 8366.
47 T. Liu, A. Barnard and Z. Tho, Residual Decomposition and Statistical Analyses of DSD, PS, and MGF Data, 2023.
48 T. Liu, Impact of Nanoparticle Morphologies on Property Prediction using Explainable AI, figshare, 2025, v6, DOI: **10.6084/m9.figshare.30254461.v6**.
49 A. S. Barnard, *Nanoscale*, 2014, **6**, 9983–9990.
50 K. E. Hermann, *J. Phys.: Condens. Matter*, 2023, **36**, 045303.
51 J. Gong, R. S. Newman, M. Engel, M. Zhao, F. Bian, S. C. Glotzer and Z. Tang, *Nat. Commun.*, 2017, **8**, 14038.
52 A. S. Barnard and G. Opletal, *Nano Fut.*, 2020, **4**, 035003.
53 R. Stocks and A. S. Barnard, *J. Phys.: Condens. Matter*, 2021, **33**, 324003.
54 H. Dong, A. S. Barnard and A. J. Parker, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 015041.

