SUPPORTING INFORMATION
Impact of Atomistic or Crystallographic Descriptors for Classification of Gold Nanoparticles

Haonan Zhang[1] and Amanda S. Barnard*[1]

[1]*School of Computing, Australian National University, Acton ACT 2601, Australia*

The document contains the algorithms for iterative label spreading (ILS) (Algorithm 1), the Tukey method (Algorithm 2) and the linear correlation removal procedure (Algorithm 3). This is followed by the t-SNE maps of the distribution of clusters, and the corresponding minimum distance plot from ILS, for results using k-Means (Figures S1 and S2) and DBSCAN (Figures S3 and S4); then the results of the classification (learning curves and classification reports) obtained using the random forest classifier (Figure S5, Table S1) and logistic regression (Figure S6, Table S2). Finally a list of features in the atomic descriptor (Table S3) and crystallographic descriptor (Table S4) are provided. The correlations between the features for both the atomic and crystallographic descriptors are shown in Figure S6 and described in detail in Reference 2. The method used for the original data generation is included at the end, and is described in detail in References 4 to 10.

*Corresponding author: amanda.s.barnard@anu.edu.au

Algorithm 1 Iterative Label Spreading

Require: Original dataset $X = \{x_1, x_2, ..., x_m\}, x_i \in R^n$;
Ensure: The List of distance D; The indexes of distance order L
Initialize two empty lists L and D
Randomly choose a point from X into L, add 0 to D and remove it from X
repeat
    Compute the all distance between labeled points and unlabeled points
    Move the nearest unlabeled point to labeled points to L, add the distance to D and remove it from X
until X is empty

Algorithm 2 Tukey method

Require: Original dataset $D = \{x_1, x_2, ..., x_m\}, x_i \in R^n$; The set of original features F; The tolerance of outliers n
Ensure: The set of the outliers index O_i
Initialize I as a empty set
for each $f \in F$ **do**
    Q_1 is the first quartile of $D[f]$;
    Q_3 is the third quartile of $D[f]$;
    R_o is the range of outlier: $R_o = 1.5 * (Q_3 - Q_1)$;
    The outliers' index for feature f, I_f is where $D[f] < Q_1 - R_o$ or $D[f] > Q_3 + R_o$;
    Append I_f in I
end for
Count the frequency of indexes in O_i. If the frequency $> n$, append this index in O_i

Algorithm 3 Remove linear correlation

Require: Original dataset $D = \{x_1, x_2, ..., x_m\}, x_i \in R^n$;
Ensure: A nonlinear correlation dataset $\hat{D}$
Initialize $\hat{D}$ as a empty set
for each $x \in D$ **do**
    Append x into $\hat{D}$;
    Compute the rank r of $\hat{D}$;
    if then$r <$ the number of features in $\hat{D}$ or all entries in x are zero
        Pop x from $\hat{D}$
    end if
end for

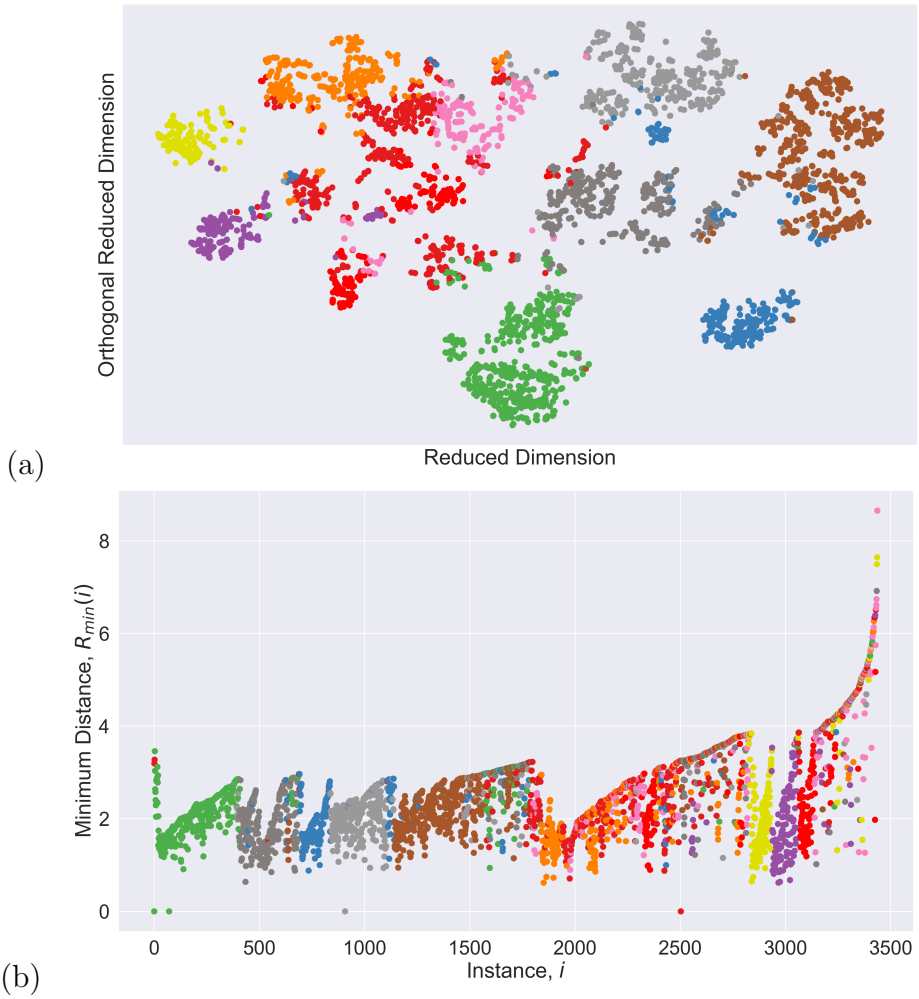


Figure S1: Results of k-Means clustering with the atomic descriptor, showing (a) the distributions of the clusters visualised with t-SNE, and (b) the minimum distance plot from ILS showing 11 clusters, and coloured by the k-Means cluster assignments.

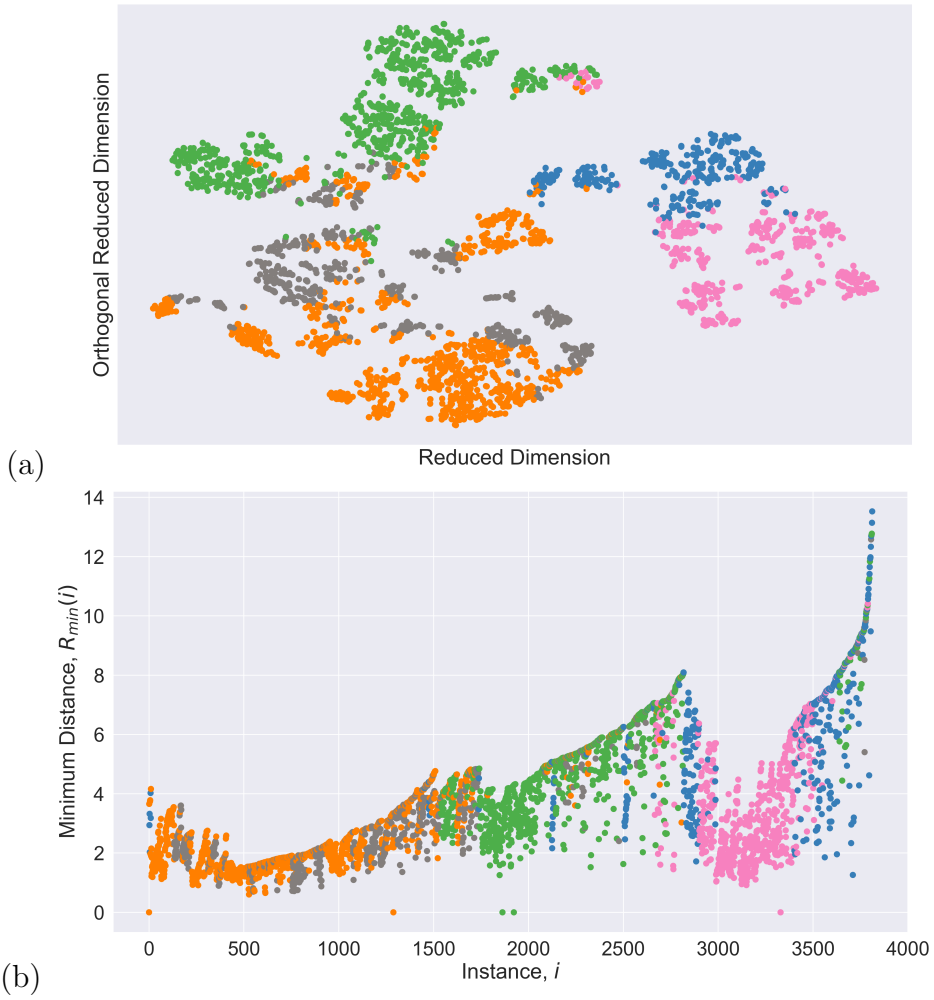


Figure S2: Results of k-Means clustering with the crystallographic descriptor, showing (a) the distributions of the clusters visualised with t-SNE, and (b) the minimum distance plot from ILS showing 11 clusters, and coloured by the k-Means cluster assignments.

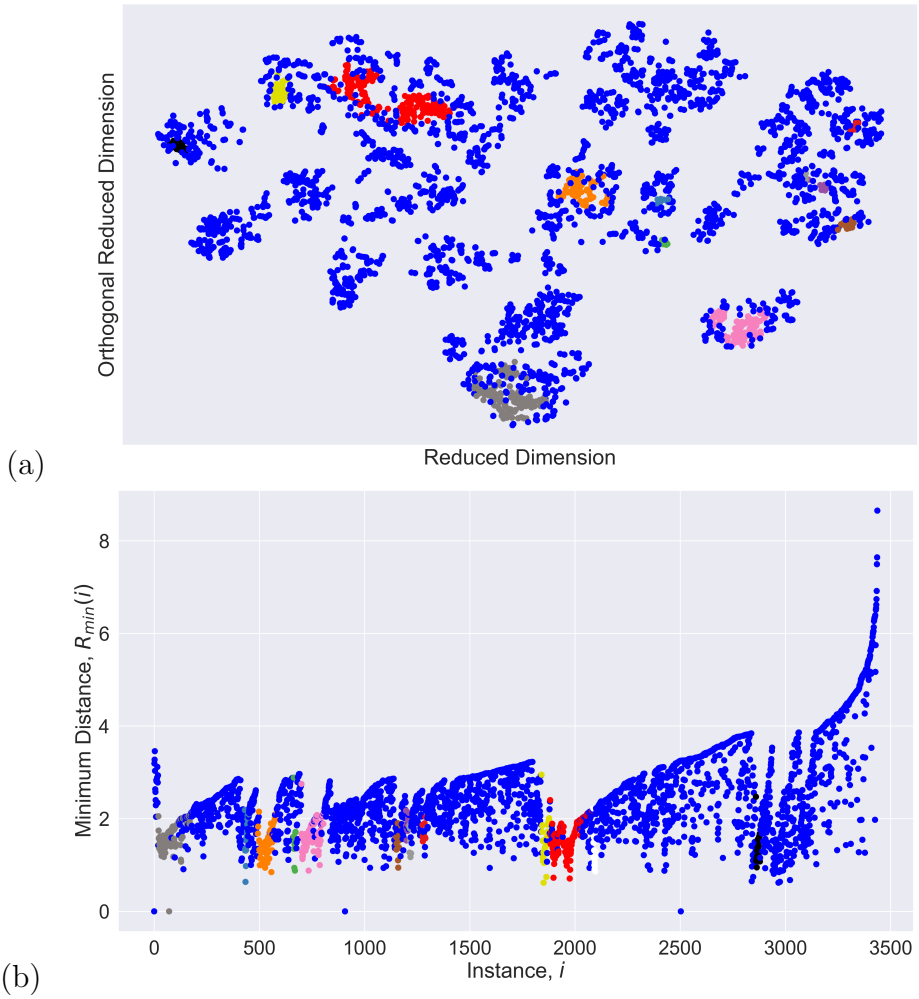


Figure S3: Results of DBSCAN clustering with the atomic descriptor, showing (a) the distributions of the clusters visualised with t-SNE, and (b) the minimum distance plot from ILS showing 11 clusters, and coloured by the DBSCAN cluster assignments.

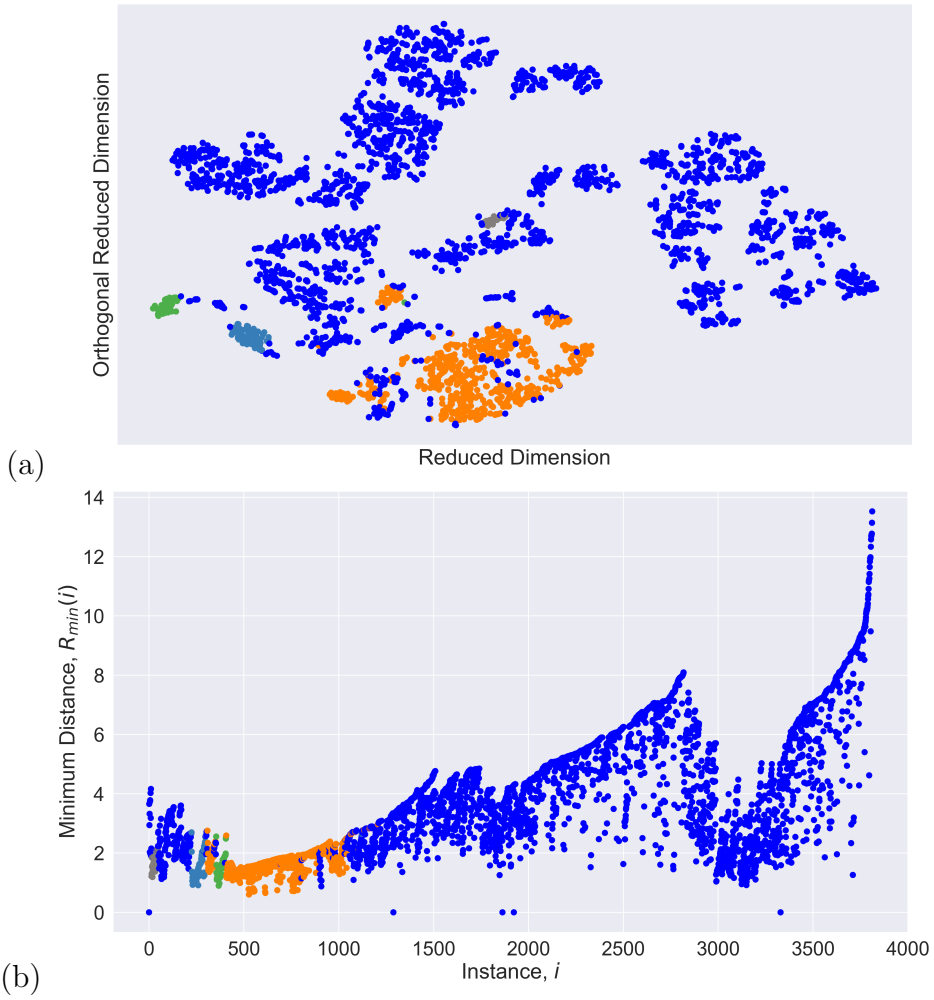
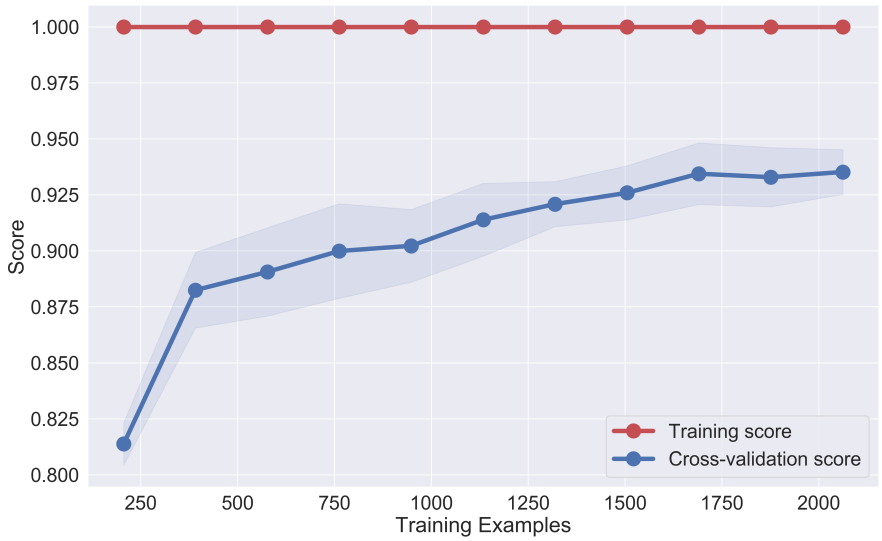


Figure S4: Results of DBSCAN clustering with the crystallographic descriptor, showing (a) the distributions of the clusters visualised with t-SNE, and (b) the minimum distance plot from ILS showing 11 clusters, and coloured by the DBSCAN cluster assignments.



(a)

(b)

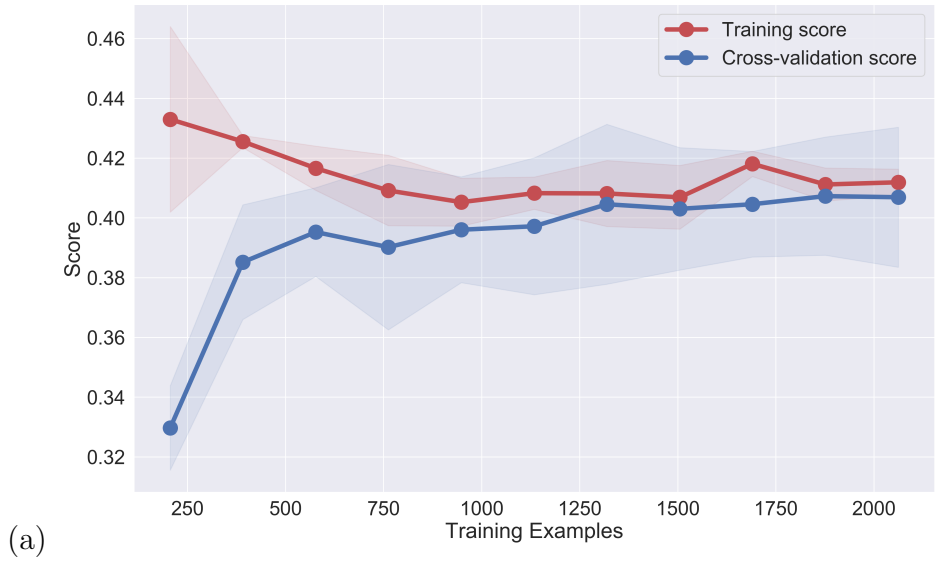
Figure S5: Results for the random forest classifier for (a) the atomic descriptor, and (b) the crystallographic descriptor, with labels obtained from the agglomerative clustering.

Table S1: Classification reports for the atomic and crystallographic descriptors from the random forest classifier using labels from agglomerative clustering.

Atomic Descriptor				Crystallographic Descriptor			
Class	**Precision**	**Recall**	**F1-score**	**Class**	**Precision**	**Recall**	**F1-score**
Class 1	0.91	0.93	0.92	Class 1	1.00	1.00	1.00
Class 2	1.00	1.00	1.00	Class 2	0.99	0.94	0.96
Class 3	0.94	0.92	0.93	Class 3	0.99	0.97	0.98
Class 4	0.96	0.97	0.97	Class 4	0.97	0.99	0.98
Class 5	0.98	0.98	0.98	Class 5	0.93	0.97	0.95
Class 6	0.98	1.00	0.99				
Class 7	0.96	0.98	0.97				
Class 8	0.89	0.95	0.92				
Class 9	1.00	0.86	0.92				
Class 10	0.96	0.96	0.96				
Class 11	0.90	0.84	0.87				

Table S2: Classification reports for the atomic and crystallographic descriptors from the logistic regression classifier using labels from agglomerative clustering.

Atomic Descriptor				Crystallographic Descriptor			
Class	**Precision**	**Recall**	**F1-score**	**Class**	**Precision**	**Recall**	**F1-score**
Class 1	0.36	0.29	0.32	Class 1	0.92	0.93	v0.93
Class 2	0.27	0.09	0.13	Class 2	0.77	0.64	0.70
Class 3	0.31	0.27	0.29	Class 3	0.74	0.87	0.80
Class 4	0.30	0.29	0.30	Class 4	0.82	0.89	0.85
Class 5	0.49	0.78	0.60	Class 5	0.55	0.34	0.42
Class 6	0.00	0.00	0.00				
Class 7	0.45	0.46	0.46				
Class 8	0.29	0.52	0.37				
Class 9	0.00	0.00	0.00				
Class 10	0.71	0.71	0.71				
Class 11	0.45	0.49	0.47				



(a)

(b)

Figure S6: Results for the logistic regression classifier for (a) the atomic descriptor, and (b) the crystallographic descriptor, with labels obtained from the agglomerative clustering. Note that the reduction in the training score with respect to the number of training instances is due to the complicity of the training set increasing with size, but the complexity of the regression remaining unchanged, results in a minor increase in under-fitting.

Table S3: List of Features in the Atomic Descriptor

Feature	Description
ID	Unique identifier, filename ID.xyz
Au_bulk	Fraction of bulk atoms
Au_surface	Fraction of surface atoms
R_min	Nanoparticle radius minimum, Å
R_max	Nanoparticle radius maximum, Å
R_diff	Nanoparticle radius minimum, Å
R_avg	Nanoparticle radius average, Å
R_std	Nanoparticle radius standard deviation, Å
R_skew	Nanoparticle radius skewness, Å
R_kurt	Nanoparticle radius kurtosis, Å
S_100	Fraction of atoms located on (100) surfaces
S_111	Fraction of atoms located on (111) surfaces
S_110	Fraction of atoms located on (110) surfaces
S_311	Fraction of atoms located on (311) surfaces
Au_coord_total	Order parameters, Average coordination number of all atoms
Au_bulk_coord	Coordination statistics, Average coordination number of all bulk atoms
Au_surface_coord	Coordination statistics, Average coordination number of all surface atoms
TCN_1	Coordination statistics, Fraction of atoms with coordination number 1
TCN_2	Coordination statistics, Fraction of atoms with coordination number 2
TCN_3	Coordination statistics, Fraction of atoms with coordination number 3
TCN_4	Coordination statistics, Fraction of atoms with coordination number 4
TCN_5	Coordination statistics, Fraction of atoms with coordination number 5
TCN_6	Coordination statistics, Fraction of atoms with coordination number 6
TCN_7	Coordination statistics, Fraction of atoms with coordination number 7
TCN_8	Coordination statistics, Fraction of atoms with coordination number 8
TCN_9	Coordination statistics, Fraction of atoms with coordination number 9
TCN_10	Coordination statistics, Fraction of atoms with coordination number 10
TCN_11	Coordination statistics, Fraction of atoms with coordination number 11
TCN_12	Coordination statistics, Fraction of atoms with coordination number 12
TCN_13	Coordination statistics, Fraction of atoms with coordination number 13
TCN_14	Coordination statistics, Fraction of atoms with coordination number 14
TCN_15	Coordination statistics, Fraction of atoms with coordination number 15
TCN_16	Coordination statistics, Fraction of atoms with coordination number 16
BCN_1	Coordination statistics, Fraction of bulk atoms with coordination number 1
BCN_2	Coordination statistics, Fraction of bulk atoms with coordination number 2
BCN_3	Coordination statistics, Fraction of bulk atoms with coordination number 3
BCN_4	Coordination statistics, Fraction of bulk atoms with coordination number 4
BCN_5	Coordination statistics, Fraction of bulk atoms with coordination number 5
BCN_6	Coordination statistics, Fraction of bulk atoms with coordination number 6
BCN_7	Coordination statistics, Fraction of bulk atoms with coordination number 7
BCN_8	Coordination statistics, Fraction of bulk atoms with coordination number 8
BCN_9	Coordination statistics, Fraction of bulk atoms with coordination number 9
BCN_10	Coordination statistics, Fraction of bulk atoms with coordination number 10
BCN_11	Coordination statistics, Fraction of bulk atoms with coordination number 11
BCN_12	Coordination statistics, Fraction of bulk atoms with coordination number 12
BCN_13	Coordination statistics, Fraction of bulk atoms with coordination number 13

Continued on next page

Feature	Description
BCN_14	Coordination statistics, Fraction of bulk atoms with coordination number 14
BCN_15	Coordination statistics, Fraction of bulk atoms with coordination number 15
BCN_16	Coordination statistics, Fraction of bulk atoms with coordination number 16
SCN_1	Coordination statistics, Fraction of surface atoms with coordination number 1
SCN_2	Coordination statistics, Fraction of surface atoms with coordination number 2
SCN_3	Coordination statistics, Fraction of surface atoms with coordination number 3
SCN_4	Coordination statistics, Fraction of surface atoms with coordination number 4
SCN_5	Coordination statistics, Fraction of surface atoms with coordination number 5
SCN_6	Coordination statistics, Fraction of surface atoms with coordination number 6
SCN_7	Coordination statistics, Fraction of surface atoms with coordination number 7
SCN_8	Coordination statistics, Fraction of surface atoms with coordination number 8
SCN_9	Coordination statistics, Fraction of surface atoms with coordination number 9
SCN_10	Coordination statistics, Fraction of surface atoms with coordination number 10
SCN_11	Coordination statistics, Fraction of surface atoms with coordination number 11
SCN_12	Coordination statistics, Fraction of surface atoms with coordination number 12
SCN_13	Coordination statistics, Fraction of surface atoms with coordination number 13
SCN_14	Coordination statistics, Fraction of surface atoms with coordination number 14
SCN_15	Coordination statistics, Fraction of surface atoms with coordination number 15
SCN_16	Coordination statistics, Fraction of surface atoms with coordination number 16

Table S4: List of Features used in the Crystallographic Descriptor

Feature	Description
ID	Unique identifier, filename ID.xyz
Au_bulk	Fraction of bulk atoms
Au_surface	Fraction of surface atoms
R_min	Nanoparticle radius minimum, Å
R_max	Nanoparticle radius maximum, Å
R_avg	Nanoparticle radius average, Å
R_std	Nanoparticle radius standard deviation, Å
R_skew	Nanoparticle radius skewness, Å
R_kurt	Nanoparticle radius kurtosis, Å
Au-Au_bonds	Bonding statistics, Average bond length, Å
Au-Au_stdev	Bonding statistics, Standard Deviation of the bond length, Å
Au-Au-Au_angle	Bonding statistics, Average bond angle, Degrees
Au-Au-Au_stdev	Bonding statistics, Standard deviation of the bond angle, Degrees
FCC	Lattice statistics, Fraction of atoms in face centred cubic (fcc) lattice
HCP	Lattice statistics, Fraction of atoms in hexagonal closed packed (hcp) lattice
ICOS	Lattice statistics, Fraction of atoms in icosahedral lattice
DECA	Lattice statistics, Fraction of atoms in decahedral lattice
q6q6_total	Order parameters, Average spherical harmonic (q6.q6 >0.7) for all atoms
q6q6_bulk	Order parameters, Average spherical harmonic (q6.q6 >0.7) for all bulk atoms
q6q6_surf	Order parameters, Average spherical harmonic (q6.q6 >0.7) for all surface atoms
q6q6_T0	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 0
q6q6_T1	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 1
q6q6_T2	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 2
q6q6_T3	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 3
q6q6_T4	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 4
q6q6_T5	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 5
q6q6_T6	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 6
q6q6_T7	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 7
q6q6_T8	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 8
q6q6_T9	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 9
q6q6_T10	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 10
q6q6_T11	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 11
q6q6_T12	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 12
q6q6_T13	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 13
q6q6_T14	Order parameters, Fraction of atoms with spherical harmonic (q6.q6 >0.7) of 14
q6q6_B0	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 0
q6q6_B1	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 1
q6q6_B2	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 2
q6q6_B3	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 3
q6q6_B4	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 4
q6q6_B5	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 5
q6q6_B6	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 6
q6q6_B7	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 7
q6q6_B8	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 8
q6q6_B9	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 9
q6q6_B10	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 10

Continued on next page

Feature	Description
q6q6_B11	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 11
q6q6_B12	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 12
q6q6_B13	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 13
q6q6_B14	Order parameters, Fraction of bulk atoms with spherical harmonic (q6.q6 >0.7) of 14
q6q6_S0	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 0
q6q6_S1	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 1
q6q6_S2	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 2
q6q6_S3	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 3
q6q6_S4	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 4
q6q6_S5	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 5
q6q6_S6	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 6
q6q6_S7	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 7
q6q6_S8	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 8
q6q6_S9	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 9
q6q6_S10	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 10
q6q6_S11	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 11
q6q6_S12	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 12
q6q6_S13	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 13
q6q6_S14	Order parameters, Fraction of surface atoms with spherical harmonic (q6.q6 >0.7) of 14

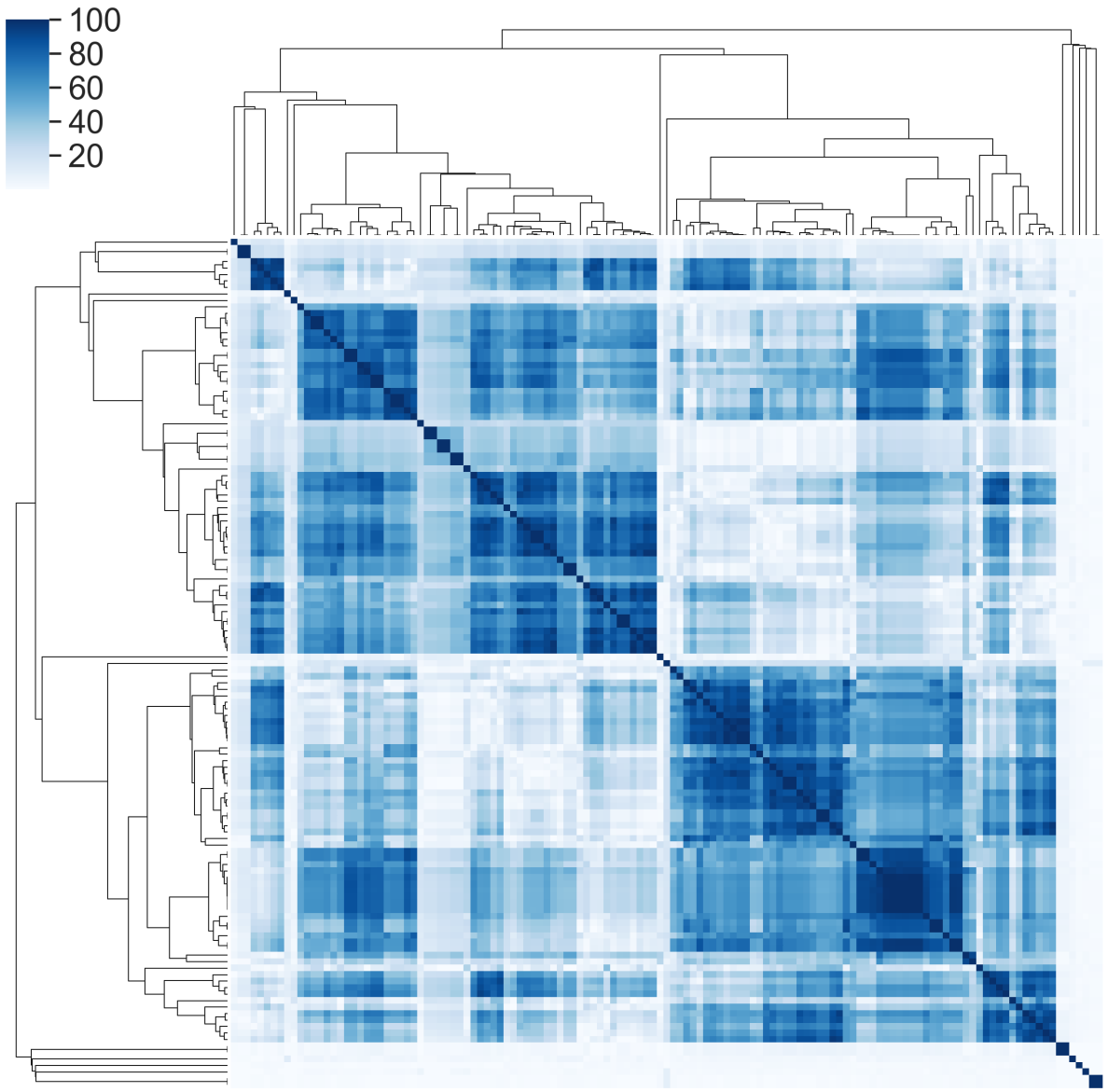


Figure S7: Correlation matrix for the entire (combined) feature space provided with Reference [1], and describe in detail in Reference [2], showing minimal correlations between the features of the atomic and crystallographic descriptors.

Original Data Generation

As mentioned in the main text, the data used in this study was not generated as part of this study. Originally these nanoparticles were generated using a combination of traditional molecular dynamics (MD) simulations and simple statistical data processing, to ensure the samples are physically realistic, thermodynamically stable and characteristic of nanoparticles observed experimentally, but guaranteed to be structurally unique.

The MD simulations of the growth process were originally performed by using the LAMMPS code [3], using embedded atom method (EAM) to accurately describe Au–Au atomic interactions. Under this scheme the total electronic density in a metal is approximated by a linear superposition of contributions from individual atoms, so the electron density in the vicinity of each atom can be expressed as the sum of the contributions given by the atom in question, plus the electron density of all neighboring atoms. In this way, for a system consisting of n atoms, the expression to calculate the energy could be written as:

$$U_{tot} = \sum_{i}^{n} G_i(\rho i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1 \neq i}^{n} \phi_{ij}(R_{ij}) \tag{1}$$

where the first term is the embedded energy and the second term represent short-range pair potential. $G_i(\rho i)$ represents the energy needed to embed an atom i in an electronic density given by (ρ). (ρi) is the electronic density pattern in atom i due to all the other atoms in the system and is defined as follows

$$\rho i = \sum_{j=1 \neq i}^{n} \rho_j^a(R_{ij}), \tag{2}$$

where ρ_j^a is the atomic electronic density with spherical distribution, contributed by the atom j. The last term in equation 1, ϕ_{ij}, represents the pair electrostatic interaction between ionic cores and is calculated using the Coulomb law,

$$\phi_{ij} = \frac{1}{4\pi\epsilon_0} \frac{Q_i(r)Q_j(r)}{r_{ij}} \tag{3}$$

where Q represents the effective charge.

In their raw form trajectories from MD simulations are not suitable for statistical analysis or machine learning, since structures from temporally adjacent time-steps may or may not be significantly different (such as those occupying a local minima before a transition state barrier could be breached). Failure to eliminate redundant structures, or those that are statistically indistinguishable based on their energy or coordinate geometries, can result in an over representation of certain types of structures which are an artefact of the simulation and not meaningful. For this reason it is important to process the data to extract structures from the trajectory when they become statistically different from the previous extraction. During this process unbound atoms, smaller clusters or nanoparticles formed *via* secondary nucleation may also be removed, to restrict the ensemble to primary nanoparticles at each extraction point. This simulated a kinetically-limited formation process. In addition to this a set of ideal polyhedron a a variety of sizes was relaxed at a set of finite temperatures to simulate a thermodynamically-limited formation process, and the data sets combined. In this way an ensemble is created that is suitable for informatics. This technique has been used to generate data sets of other metallic nanoparticles as well [5, 4, 6, 7, 8, 9, 10].

References

[1] A. Barnard, G. Opletal, *Gold Nanoparticle Data Set, v1*, CSIRO Data Collection, **2019**, `https://doi.org/10.25919/5d395ef9a4291`

[2] Barnard, A. S.; Opletal, G. Selecting Machine Learning Models for Metallic Nanoparticles. *Nano Futures* 2020, **4**, 035003–035014.

[3] `http://lammps.sandia.gov.`

[4] A.J. Parker, B. Motevalli, G. Opletal, A. S. Barnard, The pure and representative types of disordered platinum nanoparticles from machine learning. *Nanotech.* 2021, **32**, 095404.

[5] A. S. Barnard, G. Opletal, Selecting machine learning models for metallic nanoparticles. *Nano Futures* 2020, **4**, 035003

[6] A. J. Parker, G. Opletal, A.S. Barnard, Classification of platinum nanoparticle catalysts using machine learning. *J. Appl. Phys.* 2020, **128**, 014301.

[7] B. Sun, H. Barron, G. Opletal, A. S. Barnard, From process to properties: Correlating synthesis conditions and structural disorder of platinum nanocatalysts. *J. Phys. Chem. C*, 2018, **122**, 28085–28093

[8] B. Sun, H. Barron, B. Wells, G. Opletal, A.S. Barnard, Correlating anisotropy and disorder with the surface structure of platinum nanoparticles. *Nanoscale*, 2018, **10** 20393–20404

[9] H. Barron, G. Opletal, R.D. Tilley, A.S. Barnard, Predicting the role seed morphology in the evolution of anisotropic nanocatalysts. *Nanoscale*, 2017, **9**, 1502–1510

[10] H. Barron, G. Opletal, R.D. Tilley, A.S. Barnard, Dynamic evolution of specific catalytic sites on Pt nanoparticles. *Catal. Sci. Technol.* 2016, **6**, 144–151