SUPPORTING INFORMATION
Unsupervised Pattern Recognition on the Surface of Simulated Metal Nanoparticles for Catalytic Applications

Jonathan Y.C. Ting [*1], George Opletal[2], and Amanda S. Barnard [†1]

[1]*School of Computing, Australian National University, Acton 2601, Australia*
[2]*Data61, Commonwealth Scientific and Industrial Research Organisation, 3008 Melbourne, VIC, Australia*

The document contains the following sections:

- Section S1 provides further details regarding the descriptors used in this work, including the utility scores used for the removal of features from highly correlated pairs, and the software programs and parameters used to generate the features.
- Section S2 describes the dimensionality reduction methods and parameters used.
- Section S3 outlines the methodology for the identification of the cluster centres in iterative label spreading (ILS) minimum distance (R_{min}) plots, and for the automatic identification of the peaks that separate the clusters.
- Section S4 explains the internal evaluation metrics for the clustering results.
- Section S5 contains details regarding the reference activity maps and their normalisation.
- Section S6 provides details regarding the computation of the domain relevant evaluation metrics.
- Section S7 shows the correlation matrices for all nanoparticle data sets.
- Section S8 shows the cumulative explained variance plots for all nanoparticle data sets.
- Section S9 demonstrates the ability of the algorithm to distinguish between the bulk and surface atoms of ordered and disordered nanoparticles.
- Section S10 shows the illustration of the ability of the clustering pipeline to distinguish subsurface structures.
- Section S11 shows some minimum distance plots for the verification of final clustering results.
- Section S12 demonstrates the impact of the feature space on the clustering results.

[*]Corresponding author: jonathan.ting@anu.edu.au
[†]Corresponding author: amanda.s.barnard@anu.edu.au

- Section S13 illustrates the ability of the algorithm to identify peaks where human eyes might fail.
- Section S14 shows the profiles of different sets of features for all clusters identified from the surface atoms of an ordered and a disordered nanoparticle.
- Section S15 tabulates the internal evaluation scores for the clustering results.
- Section S16 shows the comparisons between the catalytic weighting profiles and the surface cluster GCN distributions for the chosen ordered and disordered nanoparticles.

S1 Descriptors

The raw data for each structure is in the form of 3D Cartesian coordinates. Feature extraction and engineering are used to make them more amenable to clustering algorithms. The former is defined as the process of transforming the raw data into more comprehensive features, while feature engineering involves generation of new features from existing information. This is done at the atomistic level in this study, where potentially useful structural features are computed for each atom of any given nanoparticle, an then grouped into descriptors, including:

a) Positional descriptor: Normalised 3D Cartesian coordinates of the atoms and their normalised radial distances from the spatial centre point of the nanoparticle. The computation of the centre of nanoparticle did not take the mass of each atom into account.

b) Geometric descriptor: Statistics (average, minimum, maximum, total number) of bond lengths, bond angles involving only first nearest neighbours, bond angles involving second nearest neighbours, and bond torsions (taking second nearest neighbours into account), along with Ackland-Jones angular parameters[1] (also known as χ parameters).

c) Steinhardt descriptor: Steinhardt's bond-orientational order parameters[2] and their averages.

d) Neighbour descriptor: Coordination number, generalised coordination number, surface coordination number, surface generalised coordination number, number of atoms with a magnitude of dot products of q_6 with itself over the threshold of 0.7.

e) Order descriptor: Centrosymmetry parameters,[3] entropy-enthalpy parameters,[4] and degeneracy degree.

Descriptors (a), (b), and (d) were generated using the *Network Characterisation Package* (NCPac),[5] (except the Ackland-Jones angular parameters), which returns various atomistic properties averaged across a given nanoparticle. The program had previously been used for feature extraction for different nanomaterials such as gold nanorods and platinum nanoparticles.[5,6] Descriptors (c) and (e), along with the Ackland-Jones angular parameters, were generated from the *Atomic Simulation Environment*[7] and *Python Structural Environment Calculator*[8] packages, with the exception being the degeneracy degree, which was extracted using *SYMMOL*.[9]

S1.1 Geometric descriptors

S1.1.1 Bond angles and torsions

An illustration of the bond angles and bond torsions computed was provided in Figure S1.

S1.1.2 Ackland-Jones parameters

The Ackland-Jones (χ) parameters characterise the local atomic environment by measuring all local angles created by each atom with its neighbours, and generating a vector from the histogram of these angles to identify crystal structures.[1] The histogram consists of the cosine values of the angles with the following bins $(-1.0, -0.945, -0.915, -0.755, -0.705, -0.195, 0.195, 0.245, 0.795, 1.0)$, with each bin corresponding to a χ_i value, where $i \in 0, 1, ..., 8$. χ_0 is excluded from our feature pools due to its exhibition of asymmetricity in supposedly symmetric atomic environments.

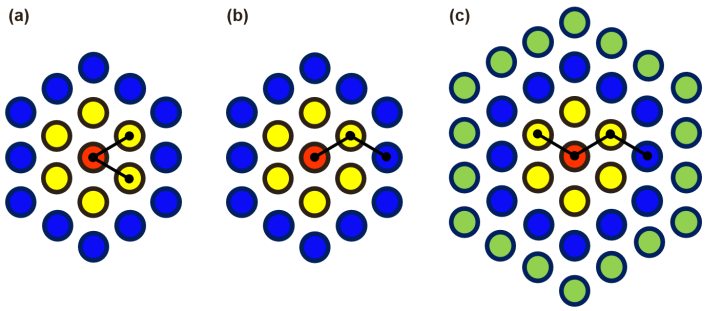


Figure S1: Illustration of (a) bond angle involving only first nearest neighbours , (b) bond angle involving second nearest neighbours, and (c) bond torsion.

S1.2 Steinhardt descriptors

S1.2.1 Steinhardt's bond-orientational order parameters

Steinhardt's bond-orientational order parameters, q_l (Equation 1), have been widely used for identification of crystallinity (q_2 and q_6), quantification of solidity or liquidity (q_6), distinction of crystal structures (q_4 and q_6), and detection of defects,[10] due to their desirable rotationally and translationally invariant properties.

$$q_l(i) = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2 \right)^{\frac{1}{2}} \tag{1}$$

$$q_{lm}(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{lm}(\mathbf{r}_{ij}) \tag{2}$$

where:
Y_{lm} = spherical harmonics of degree l and order m, $m \in [-l, +l]$, $m, l \in \mathbb{Z}$
N_i = number of first shell neighbours of atom i
$\mathbf{r}_{ij}$ = vector from atoms i to j

However, finite temperatures always cause fluctuations in the atomic positions, leading to overlapping distributions of the q_l parameters, making the determination of crystal structures difficult. The averaged Steinhardt's parameters, $\bar{q}_l$ (Equation 3), were proposed by Lechner and Dellago in 2008 to address the issue.[11] The overlap between the distributions is reduced by incorporating information from the first nearest neighbour shell, allowing easier identification of crystal structures (commonly $\bar{q}_4$ and $\bar{q}_6$).

$$\bar{q}_l(i) = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left| \frac{1}{N_i} \sum_{k=0}^{N_i} q_{lm}(k) \right|^2 \right)^{\frac{1}{2}} \tag{3}$$

S1.2.2 Neighbour list

The Steinhardt descriptors are very sensitive to the neighbour list. Three neighbour-finding algorithms were tested, namely solid-angle based nearest-neighbor,[12] Voronoi tessellation,[13] and adaptive cutoff neighbour-finding algorithm.[14] The solid-angle based nearest-neighbour algorithm equates the solid angles around every atom to 4π and solves Equation 5 iteratively to determine the cutoff radius, R, whereas

the Voronoi tessellation approach identifies the neighbours of each atom by partitioning the 3D space into regions closest to each atom and counting the number faces of the Voronoi polyhedra of the atom.[13] The latter approach is purely geometric and does not require any parameter except the specification of distance metric, which is by default Euclidean. However, both of them failed to identify the neighbours of the edge atoms correctly, as shown in Figure S2. Hence, the neighbour lists in this work are computed using the adaptive cutoff neighbour-finding scheme, which applies a padding to the average distance of a specified number of nearest neighbours of a given central atom (Equation 4).

$$r_{\text{cut}}(i) = p\left(\frac{1}{n}\sum_{j=1}^{N_{\max}} r_{ij}\right) \tag{4}$$

where:

p = multiplier to be applied to the cutoff distance as safe padding
$N_{\max}$ = maximum number of atoms to be considered, default value is 6
$r_{i,j}$ = distance between atoms i and j

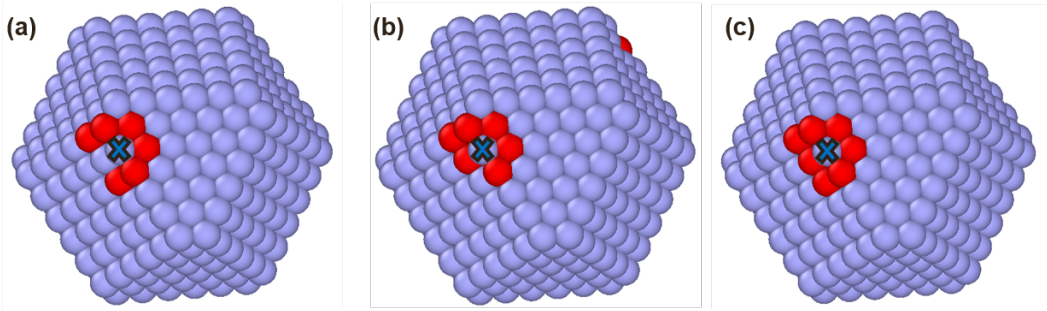


Figure S2: Illustrations of the neighbour (red) atoms identified by the (a) solid angle-based nearest neighbour, (b) Voronoi tessellation, and (c) adaptive cutoff neighbour-finding algorithms given a metal nanoparticle edge atom (marked by blue crosses).

$$R_i^{(m)} = \frac{\sum_{j=1}^{m} r_{i,j}}{m-2} < r_{i,m+1} \tag{5}$$

where:

m = number of neighbours of atom i, which is increased iteratively

S1.3 Neighbour descriptors

S1.3.1 Coordination number

The surface coordination number is obtained by applying the constraint of only including surface atoms during the counting of neighbouring atoms for the computation of coordination number.

S1.3.2 Number of bonded neighbours

As mentioned in Section S1.2, the q_6 parameter could be used to detect crystalline atoms. Atoms i and j are considered bonded if the magnitude of $\sum_{-l\leq m\leq l} q_{lm}^{i}\left(q_{lm}^{j}\right)^{*}$ for $l = 6$ exceeds a threshold of 0.7, as implemented in previous studies.[15,16] The feature $q6q6$ denotes the number of neighbouring atoms that are bonded to the atom of interest. The feature is found to be equivalent to coordination number

in ideally-shaped polyhedral nanoparticles, but provides complementary information when applied to more disordered nanoparticles.

S1.4 Order descriptors

S1.4.1 Centrosymmetry parameter

Another parameter that was introduced to identify crystal defects is the centrosymmetry parameter, CSP (Equation 6).[3] This is done by measuring the loss of symmetry (degree of inversion symmetry of an atomic environment), which requires an algorithm to identify the opposite pairs of neighbours.

For this work, the algorithm used is Greedy Edge Selection,[17] which computes a weight, $w_{ij} = |\mathbf{r}_i + \mathbf{r}_j|$ for all combinations of neighbour pairs around atom i. $\frac{N}{2}$ combinations with the smallest weights are used for the calculation of CSP. An alternative algorithm is Greedy Vertex Matching,[18] which orders neighbouring atoms by their distances from the central atom, and pairs the closest neighbour with its lowest weight partner. The latter has been found to be more sensitive to perturbations and thus is not employed here.[18]

$$CSP = \sum_{i=1}^{\frac{N_{\text{bulk}}}{2}} \|\mathbf{r}_i + \mathbf{r}_{i+\frac{N_{\text{bulk}}}{2}}\|^2 \tag{6}$$

where:

N_{bulk} = total number of neighbouring atoms, 12 for face-centred cubic and hexagonal close-packed, 8 for body-centred cubic

$\mathbf{r}_i$ and $\mathbf{r}_{i+\frac{N_{\text{bulk}}}{2}}$ = vectors from central atom to two opposite pairs of neighbours

S1.4.2 Entropy parameter

The entropy parameter, s_s (Equation 7), is a relatively nascent concept proposed by Piaggi to identify defects and distinguish between solid and liquid.[4] The averaged entropy parameter, $\bar{s}_s^i$ (Equation 9) is calculated using a simple average.

$$s_s^i = -2\pi\kappa_B\rho \int_0^{r_m} \left[g_m^i(r) \ln\left(g_m^i(r)\right) - g_m^i(r) + 1\right] r^2 dr \tag{7}$$

where:

r_m = upper bound of integration

g_m^i = radial distribution function centred on atom i

$$g_m^i(r) = \frac{1}{4\pi\rho r^2} \sum_j \frac{1}{\sqrt{2\pi\sigma^2}} e^{-frac(r-r_{ij})^2 2\sigma^2} \tag{8}$$

where:

r_{ij} = interatomic distance between atoms i and j

σ = broadening parameter

$$s_s^i = \frac{\sum_{j=1}^{N_i} s_s^j + s_s^i}{N+1} \tag{9}$$

S1.4.3 Symmetricity parameter

The neighbour list used to compute the Steinhardt descriptors is also used to calculate the degree of degeneracy, which is defined on the basis of the similarity of the principal moments of inertia.[9]

S1.5 Feature description

The descriptions of all features, and their corresponding utility scores used for the removal of features from highly correlated pairs, are tabulated in Table S1.

Acronym	Description	Utility score
x,y,z	Atomic coordinates in the x, y, and z axes.	10
rad	Normalised radial distance from nanoparticle centre.	0
blavg	Average bond length ($\mathring{A}$).	2
blmax	Maximum bond length ($\mathring{A}$).	8
blmin	Minimum bond length ($\mathring{A}$).	8
blnum	Number of bonds.	3
ba1avg	Average bond angles involving only first nearest neighbours (°).	2
ba1max	Maximum bond angles involving only first nearest neighbours (°).	8
ba1min	Minimum bond angles involving only first nearest neighbours (°).	8
ba1num	Number of bond angles involving only first nearest neighbours.	3
ba2avg	Average bond angles involving second nearest neighbours (°).	2
ba2max	Maximum bond angles involving second nearest neighbours (°).	8
ba2min	Minimum bond angles involving second nearest neighbours (°).	8
ba2num	Number of bond angles involving second nearest neighbours.	3
btposavg	Average positive bond torsion (°).	2
btposmax	Maximum positive bond torsion (°).	8
btposmin	Minimum positive bond torsion (°).	8
btposnum	Number of positive bond torsions.	3
btnegavg	Average negative bond torsion (°).	2
btnegmax	Maximum negative bond torsion (°).	8
btnegmin	Minimum negative bond torsion (°).	8
btnegnum	Number of negative bond torsions.	3
chiX	χ_{X-1} Ackland-Jones parameter.	6
q2	q_2 Steinhardt parameter.	5.7
q4	q_4 Steinhardt parameter.	5.6
q6	q_6 Steinhardt parameter.	5.5
q8	q_8 Steinhardt parameter.	5.7
q10	q_{10} Steinhardt parameter.	5.7
q12	q_{12} Steinhardt parameter.	5.7
q2avg	Averaged q_2 Steinhardt parameter.	5.2
q4avg	Averaged q_4 Steinhardt parameter.	5.1
q6avg	Averaged q_6 Steinhardt parameter.	5
q8avg	Averaged q_8 Steinhardt parameter.	5.2
q10avg	Averaged q_{10} Steinhardt parameter.	5.2
q12avg	Averaged q_{12} Steinhardt parameter.	5.2
cn	Coordination number.	1
gcn	Generalised coordination number.	0
scn	Surface coordination number.	3
q6q6	Number of bonded neighbours ($q_6(i) \cdot q_6(j)$ exceeding 0.7).	2
centParam	Centrosymmetry parameter.	4
entroParam	Entropy parameter.	5
entroAvgParam	Averaged entropy parameter.	5.5
degenDeg	Degeneracy degree.	6

Table S1: Description of features used in this work and their corresponding utility scores (in the range $[0, 10]$). Lower scores correspond to higher utility due to greater ease of interpretation of higher relevance to catalysis.

S1.6 Software parameters

S1.6.1 Network Characterisation Package

Network Characterisation Package (NCPac)[19] returns various atomic structural properties averaged across a given nanoparticle. For any given conformation, NCPac identifies the first nearest neighbours of all atoms and the surface atoms before computing the structural properties of all atoms and averaging over them. The nearest neighbours are identified using cutoff distance specified by the users, while the surface atoms are identified using a cone angle algorithm, depicted in Algorithm 1. Table S2 lists the input parameters used to generate the atomistic features via NCPac.

Algorithm 1 Pseudocode for determining whether a given atom is a surface atom.

Input:
n : number of points to be generated on the spherical surface of the atom
$\overrightarrow{i}$: atomic coordinates of the given atom
$J : \{\overrightarrow{j_1}, \overrightarrow{j_2}, ..., \overrightarrow{j_m}\}$ list of atomic coordinates of nearest neighbours of the given atom
θ_{thresh} : threshold of cone angle
Output:
$isAtom$: whether the atom is a surface atom

procedure CONEANGLESURF(n, $\overrightarrow{i}, J, \theta_{thresh}$)
    Generate n equidistant spherical mesh of points $K : \{\overrightarrow{k_1}, \overrightarrow{k_2}, ..., \overrightarrow{k_n}\}$ at unit radius around the atom using the Rakhmanov algorithm[20].
    $\theta_{min} \leftarrow 180$
    for $\overrightarrow{j}$ $in J$ **do**
        for $\overrightarrow{k}$ $in K$ **do**
            $v_1 \leftarrow \overrightarrow{i} - \overrightarrow{j}$
            $v_2 \leftarrow \overrightarrow{i} - \overrightarrow{k}$
            $\theta \leftarrow$ angle between v_1 and v_2 ▷ Angle computed using dot product
            if $\theta < \theta_{min}$ **then**
                $\theta_{min} \leftarrow \theta$
            end if
        end for
    end for
    if $\theta_{min} > \theta_{thresh}$ **then**
        $isAtom \leftarrow 1.$ ▷ Atom is a surface atom.
    else
        $isAtom \leftarrow 0.$ ▷ Atom is not a surface atom.
    end if
end procedure

S1.6.2 Python Structural Environment Calculator

The nearest neighbours lists are computed using the function and parameters *find_neighbors*(*method*='cutoff', *cutoff*='adaptive', *threshold*=2, *filter*=None, *padding*=1.2, *nlimit*=6). The following functions and parameters were used to calculate the:

Parameter	Description	Value
in_xyz_prec	Decimal places in coordinates.	4
in_bound_dist	Buffer distance from box boundaries in each axis.	10.0
in_cutoff(i,j)	Nearest neighbour cutoff matrix between elements i and j.	Pd 3.3
in_xl, *in_yl*, *in_zl*	Cell lengths in x, y, z directions ($\AA$).	30.0, 30.0, 30.0
in_density	Reduced density.	0.05
in_delr	Spacing interval for radial distribution function ($\AA$).	0.08
in_gr_points	Number of points for radial distribution function.	120
in_sq_points	Number of $S(q)$ points (must be $2^n + 1$, *i.e.* 65, 129, 257, ...).	513
in_cluster_flag	Whether to filter out small atom clusters.	0
in_surf_flag	Whether to find and analyse the surface layer.	1
in_cone_angle	Cone angle beyond which defines surface particles (°).	45
in_surf_points	Number of points in spherical point distribution.	3000
in_chain_flag	Whether to conduct chain analysis.	0
in_q6order_flag	Whether to conduct q_6 order analysis.	1
in_q6order_dotmin	Minimum $q_6(i) \cdot q_6(j)$ threshold for atoms i and j to be classified as similarly bonded.	0.7
in_sc_flag	Whether to conduct signature cells analysis.	0
in_SU_flag	Whether to conduct structural unit analysis.	0
in_fradim_flag	Whether to conduct fractal dimension analysis.	0
in_lindem_flag	Whether to calculate Lindemann index.	0

Table S2: Input parameters for Network Characterisation Package.

1. χ parameters: *calculate_chiparams*(*angles*=False)

2. Steinhardt's parameters: *calculate_q*(q=[2, 4, 6, 8, 10, 12], *averaged*=True, *only_averaged*=False, *condition*=None, *clear_condition*=False)

3. centrosymmetry parameter: *calculate_centrosymmetry*(*nmax*=CN, *get_vals*=True), where CN is the coordination number of the given atom.

4. entropy parameters: *calculate_entropy*(*rm*=1.4×LC, *sigma*=0.2, *rstart*=0.001, *h*=0.001, *local*=False, M=12, N=6, *ra*=None, *averaged*=True, *switching_function*=False), where LC is the lattice constant of face-centred cubic palladium (3.89).

S1.6.3 SYMMOL

The source code for SYMMOL is freely available at `https://github.com/fxcoudert/symmol`. As orthogonal coordinates are used, the cell parameters in the first line of the input file for the SYMMOL program are set to "1 1 1 90 90 90". The *INDWGH* and *INDTOL* parameters are set to 2 and 1 respectively, to calculate the moment of inertia with weights of 1.0, using *DCM* as the acceptance threshold for the symmetry element of a given molecular group. *DCM* is initialised as 0.1 and increased by 0.1 until a symmetry element is found, or *DCM* reaches 10.0.

S2 Dimensionality reduction

S2.1 Principal component analysis

Principal component analysis is used to reduce the dimensionality of the feature space, while preserving as much variance in data as possible. This is achieved by projecting each data sample onto a specified number of principal components of high-dimensional data, which are directions that maximise the variance of the projected data. In our case, the number is determined such that 99% of the variance in data is preserved. The *scikit-learn* package implementation is employed in this work, with the parameters being *n_components*=0.99, *whiten*=False, *svd_solver*='auto', *tol*=0.0, *iterated_power*='auto', *random_state*=42.

S2.2 *t*-Stochastic Neighbour Embedding

To visualise the high-dimensional data in two dimensional figures, we used *t*-distributed stochastic neighbour embedding (*t*-SNE) to learn the representation of the data in 2D manifold.[21] *t*-SNE is chosen over alternative approaches such as uniform manifold approximation and projection (UMAP), which was regarded as superior to the former in terms of preservation of the global structure of data and run time performance, due to its tuning flexibility for the display of local structures in data. Moreover, it was recently found that the *t*-SNE performs as well as UMAP when informative initialisation (such as principal component analysis initialisation) is used.[22]

The *t*-SNE algorithm attempts to embed high-dimensional data into low-dimensional space such that similar samples in original feature space are close to each other in the embedded space, and vice versa. The algorithm first constructs a probability distribution over pairs of high-dimensional samples where the probability is proportional to their similarity. A similar probability distribution is then defined over the samples in the low-dimensional space. The Kullback-Leibler divergence between the two distributions is minimised with respect to the locations of the samples in the map. The *scikit-learn* package implementation is also used, with the parameters being *n_components*=2, *perplexity*=30.0, *early_exaggeration*=12.0, *learning_rate*='auto', *n_iter*=1000, *n_iter_without_progress*=300, *min_grad_norm*=1e-7, *metric*=metric, *random_state*=42, *init*='pca', *method*='barnes_hut', and *angle*=0.5.

S2.2.1 Visualisations of ordered nanoparticles data sets

Figure S3 visualises the rest of the highly ordered nanoparticles simulated at 0 K, which are omitted in the main text.

S2.2.2 Impact of perplexity values

It is commonly known that different values of the *perplexity* parameter can significantly impact the resulted projection of *t*-SNE.[21,23] The parameter dictates the balance between preserving the global and the local structure of the data, by specifying the size of the neighbourhood used for attracting points.[23] To investigate its impact on the resulting visualisations and look out for potential projection artefacts, we show the changes in the resulted projections as the *perplexity* value is varied over a wide range in Figure S4. It is observed that the bulk and surface atoms consistently form different groups, which boosts our confidence for the statement made in the main text.

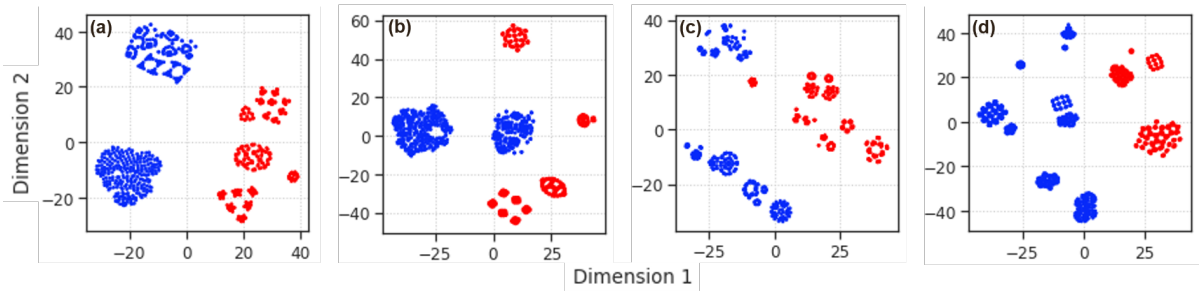


Figure S3: Mapping of the high dimensional data sets for the (a) cuboctahedron, (b) cube, (c) decahedron, and (d) icosahedron nanoparticles simulated at 0 K onto 2D manifold learnt via *t*-distributed stochastic neighbour embedding. The red and blue points correspond to surface and bulk atoms, respectively.

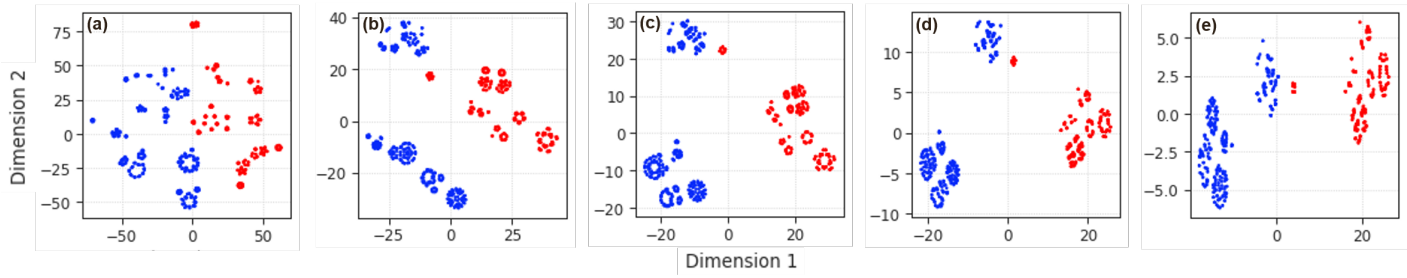


Figure S4: Mapping of the high dimensional data sets for the tetrahedron nanoparticle simulated at 0 K onto 2D manifold learnt via *t*-distributed stochastic neighbour embedding, with *perplexity* of (a) 10, (b) 30, (c) 50, (d) 100, and (e) 120. The red and blue points correspond to surface and bulk atoms, respectively.

S3 Cluster identification

S3.1 Identification of cluster centres

ILS requires one or more initialised label(s) since it is a semi-supervised clustering approach. The overall pipeline can be made unsupervised by automating the label initialisation step. ILS works the best when the initialised labels are located at the centres of any given cluster. Therefore, we have utilised an approach developed to identify the cluster centres automatically in this work, the source code of which is freely available at `https://github.com/Thea-Hsu/EILS`.

The most prominent cluster centre is first identified via a method proposed by Rodriguez,[24] which is based on the idea that cluster centres are characterised by higher density than their neighbours, and by a relatively large distance from other samples with higher densities. The sample with the highest γ score (Equation 12), which is computed based on their local density ρ (Equation 10) and minimum distance of the sample from other samples with higher density δ (Equation 11), is determined as the most prominent cluster centre, which is then solely labelled for the initial parse of ILS. The centres of each cluster identified from the initial R_{min} plot are also identified by comparing the the γ scores of the samples within the cluster. By labelling these cluster centres with different labels and rerunning ILS on all samples, another R_{min} plot that groups all samples into different clusters can then be obtained. To make sure all clusters have a size larger than a threshold (specified to be 10 in this work), the centres for clusters that are smaller than the threshold can be removed before rerunning ILS to obtain the final R_{min} plot.

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{10}$$

where $\chi(x) = 1$ if $x < 0$, otherwise $\chi(x) = 0$, d_c = cutoff distance.

$$\delta_i = \min_{j:p_j > p_i} (d_{ij}) \tag{11}$$

except for the point with highest density ($\delta_i = \max_j (d_{ij})$), where:

$$\gamma_i = \rho_i \delta_i \tag{12}$$

S3.2 Peak identifying algorithm

The clusters can be identified automatically by dividing the R_{min} plot at each peak into separate regions. We have developed a peak-finding algorithm for this purpose, as illustrated in Algorithm 2. The algorithm first identifies all local peaks by computing a peak function S (Equation 13) proposed by Palshikar,[25] which is the average of the sum of averages of the signed distances of x_i from its k left and right neighbours. The local peaks are the samples with a positive S value, and are considered as candidates for true peaks, such that:

$$S = \frac{\frac{\sum_{j=1}^{k}(x_i - x_{i-k})}{k} + \frac{\sum_{j=1}^{k}(x_i - x_{i+k})}{k}}{2} \tag{13}$$

where x is a series of samples, and k is the minimum cluster size.

The peak candidates have to satisfy three criteria to be considered a true peak. Firstly, the peak has to be a maximum within a scanning window $2k$. Secondly, the signal-to-noise ratio of the peak has

to exceed a proportion of the global maximum peak value, defined as threshold h. Lastly, the peak also needs to be sufficiently different from the other identified peaks, with a specified peak function value difference threshold t. This is set as a requirement because the ordered nanoparticles data sets are highly prone to false positives. For this work, we have set the values of k, h, and t to be 25, 0.06, and 0.01, respectively. The values here were chosen based on empirical experiments, but the parameters can be tuned to vary the sensitivity of the algorithm.

Algorithm 2 Pseudocode for the algorithm to find peaks in minimum distance plots returned by ILS clustering algorithm.

Input:
X : Minimum distance data: $x_1, ..., x_n$
k : Window size $\in \mathbb{R}$
h : Significance constant $\in [1.0, 3.0]$
t : Difference threshold $\in \mathbb{R}$
Output:
P : Peaks: p_1, ..., p_k

procedure FINDPEAKS(X, k, h, t)
  $V \leftarrow \emptyset$
  for $i = 2, ..., n$ **do**
    Compute $S(X, k)$ ▷ Equation 13
    $V \leftarrow V \cup S(X, k)$
  end for
  $P \leftarrow \emptyset$
  for $i = 2, ..., n$ **do**
    if $S(i) < h \max(S)$ **then** ▷ Not globally significant
      Skip candidate
    end if
    for $j = 1, ..., k$ **do**
      if $S(i) - S(j) < h$ **then** ▷ Not different enough
        Skip candidate
      end if
    end for
    if $i - P_{\text{prev}} < k$ **then** ▷ Multiple peaks within k
      $P \leftarrow P \cup \arg\max \{S(P_{\text{prev}}), S(i)\}$
    end if
  end for
  return P
end procedure

S4 Internal cluster evaluation

The Silhouette coefficient (Equation 14),[26] Calinski-Harabasz index (Equation 15),[27] and Davies-Bouldin index (Equation 18)[28] are used to evaluate the clustering results in the main text.

$$SC = \sum_{i=1}^{N} \frac{b\left(i\right) - a\left(\right)}{\max\left(a\left(i\right), b\left(i\right)\right)} \tag{14}$$

where:

N = total number of samples in data

a = mean distance between sample i and all other samples in the same cluster

b = mean distance between sample i and all other samples in the next nearest cluster

$$CH = \frac{tr\left(B_k\right)}{tr\left(W_k\right)} \times \frac{n_E - k}{k - 1} \tag{15}$$

where:

$tr\left(B_k\right)$ = trace of between-cluster dispersion matrix

$tr\left(W_k\right)$ = trace of within-cluster dispersion matrix

$$B_k = \sum_{q=1}^{k} n_q \left(c_q - c_E\right)\left(c_q - c_E\right)^T \tag{16}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} \left(x - c_q\right)\left(x - c_q\right)^T \tag{17}$$

where:

k = number of clusters

C_q = set of samples in cluster q

c_q = centre of cluster q

c_E = centre of data E

n_q = number of samples in cluster q

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}, \text{ where } R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{18}$$

where:

s_i = average distance between each member of cluster i and the cluster centroid

d_{ij} = distance between cluster centroids i and j

R_{ij} is chosen such that it is non-negative and symmetric.

S5 Reference activity maps

Figure S5 shows the activity maps for some reactions prior to and after normalisation, which is conducted while preserving the gradients of each equation instead of the intercepts (or equivalently, the range of the activity maps) as we are more interested in the range of GCN that are the most relevant to any given chemical reaction. It is noted that these maps were obtained using different reference electrodes. However, as the activity maps are obtained with the assumption that the GCN-activity relationship is linear, different choices of reference electrodes only result in shifting of the intercepts of the fitted equations, which is accounted for by the normalisation. Readers are directed to the corresponding publications mentioned in Section 1 in the main text for further details.

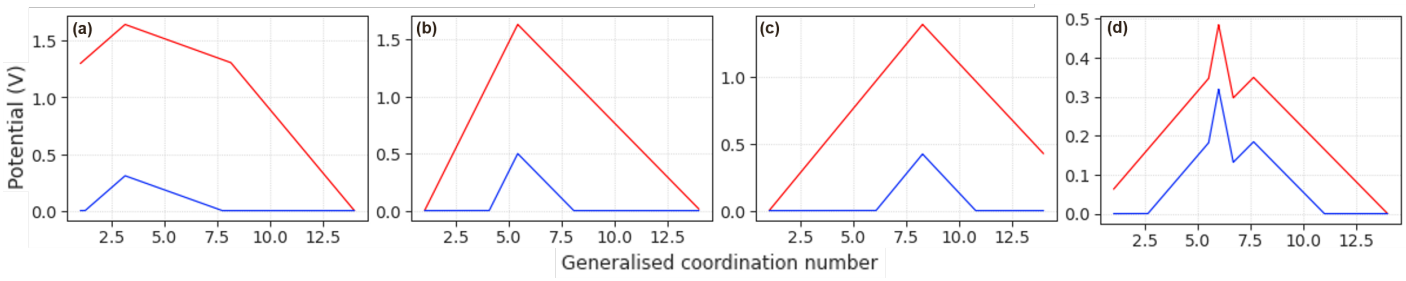


Figure S5: Reference activity maps for (a) carbon dioxide reduction, (b) carbon monoxide oxidation, (c) oxygen reduction, and (d) aliphatic ketone reduction reactions prior to and after normalisation. The red and blue lines correspond to the original and adjusted maps (so that the area under the lines sums up to 1), respectively.

S6 Domain relevant cluster evaluation

S6.1 Computation of $p(g)$

$p(g)$ is the GCN distribution of a given cluster. The distribution is computed via the fast Fourier transform based kernel density estimation (KDE) method provided by the publicly available Python package *KDEpy*. We applied Silverman's rule of thumb for bandwidth selection and used Gaussian kernels for the computation. The KDE is first fitted to the GCN values of the surface atoms of a given cluster, and then evaluated on a grid with the range of $[1, 14]$ with a spacing of 0.001.

S6.2 Computation of $q(g)$

$q(g)$ is the catalytic weighting profile obtained from the reference reaction GCN-activity map. The activity maps are composed of multiple linear equations. We list below the equations obtained from the previous publications:

Equations 19 and 20 form the activity map for oxygen reduction reaction catalysed by gold nanoparticles[29]:

$$u_{AuORR_1} = 0.128g + 0.201 \tag{19}$$

$$u_{AuORR_2} = -0.160g + 1.686 \tag{20}$$

Equations 21 and 22 form the activity map for oxygen reduction reaction catalysed by platinum nanoparticles[29]:

$$u_{PtORR_1} = 0.192g - 0.715 \tag{21}$$

$$u_{PtORR_2} = -0.169 + 2.270 \tag{22}$$

Equations 23 and 24 form the activity map for carbon monoxide oxidation reaction catalysed by platinum nanoparticles[30]:

$$u_{PtCOOR_1} = 0.3701g - 2.464 \tag{23}$$

$$u_{PtCOOR_1} = -0.1886g + 0.560 \tag{24}$$

Equations 25 and 26 form the activity map for reverse water-gas shift reaction catalysed by copper nanoparticles[31]:

$$u_{CuRWGSR_1} = 0.30833g - 0.880 \tag{25}$$

$$u_{CuRWGSR_2} = 0.1875g + 2.3225 \tag{26}$$

Equations 27 to 29 form the activity map for reverse water-gas shift reaction catalysed by copper nanoparticles[32]:

$$u_{CuCO2RR_1} = 0.163g - 1.133 \tag{27}$$

$$u_{CuCO2RR_2} = -0.067g - 0.416 \tag{28}$$

$$u_{CuCO2RR_3} = -0.223g + 0.853 \tag{29}$$

Equations 30 to 34 form the activity map for aliphatic ketone reduction reaction catalysed by platinum nanoparticles[33]:

$$u_{PtRCORRR_1} = 0.063g - 0.350 \tag{30}$$

$$u_{PtRCORRR_2} = 0.290g - 1.600 \tag{31}$$

$$u_{PtRCORRR_3} = -0.270g + 1.750 \tag{32}$$

$$u_{PtRCORRR_4} = 0.055g - 0.420 \tag{33}$$

$$u_{PtRCORRR_5} = -0.055g + 0.420 \tag{34}$$

The algorithm to normalise a given GCN-activity map is described in Algorithm 3. For each reaction of interest, the potential values within a range of GCN values of interest are first computed from the linear equations forming the activity map. The $y = 0$ line is then moved while keeping the gradients of the linear equations unchanged until the total area under the composite function is approximately 1. While Simpson's composite integration and Romberg's method are known for their capability of returning higher accuracy for area under curves, the simpler trapezoidal method is used here as the composite functions are comprised of only straight lines.[34] Readers are referred to the notebooks made available at `https://github.com/Jon-Ting/metal-nanoparticle-surface-atom-labelling` for the exact code used.

S6.3 Computation of $O(p, g)$

$O(p, q)$ is the overlapped GCN range between $p(g)$ and $q(g)$. The algorithm to compute the overlapping range is shown in Algorithm 4.

Algorithm 3 Pseudocode for the algorithm to normalise a given generalised coordination numbers-activity map into a catalytic weighting profile for a particular chemical reaction.

Input:

$G : \{g_1, ..., g_n\}$ Generalised coordination numbers in the range of interest
$M : \{m_1, ..., m_k\}$ Gradients of linear equations for the given chemical reaction
$C : \{c_1, ..., c_k\}$ Intercepts of linear equations for the given reaction
$A_{diff_{thresh}}$: Area difference threshold

Output:

$U_{adjusted} : \{u_1, ..., u_n\}$ Catalytic weighting profile

procedure qNormalise(G, M, C, A)
  $S : \{s_1, ..., s_{k-1}\} \leftarrow g$ values at the intersection points of each pair of consecutive linear equations
  $i \leftarrow 0$
  $U \leftarrow \emptyset$
  $m_l, m_r \leftarrow M_i, M_{i+1}$ ▷ Initialise left and right gradients and intercepts
  $c_l, c_r \leftarrow C_i, C_{i+1}$
  for $g \in G$ **do**
    if $(g > S_i)$ & $(i < k - 1)$ **then** ▷ Whenever an intersection is encountered
      $i \leftarrow i + 1$
      $m_l, m_r \leftarrow M_i, M_{i+1}$ ▷ Update gradients and intercepts
      $c_l, c_r \leftarrow C_i, C_{i+1}$
    end if
    if $m_r < m_l$ **then** ▷ Take minima of both lines if right slope is smaller in value
      $U \leftarrow U \cup \min(m_l g + c_l, m_r g + c_r)$
    else ▷ And vice versa
      $U \leftarrow U \cup \max(m_l g + c_l, m_r g + c_r)$
    end if
  end for
  $U \leftarrow U - \min(U)$ ▷ Ensure all values are positive
  $A_{upper}, A_{lower} \leftarrow$ current area under g curve, 0.0 ▷ Area computed using trapezoidal rule
  $v_{upper}, v_{lower} \leftarrow 0, \max(U)$
  while $A_{diff} < A_{diff_{thresh}}$ **do** ▷ Until area converges to targeted value
    $r_{diff} \leftarrow \frac{1-\sqrt{A_{lower}}}{\sqrt{A_{upper}}-\sqrt{A_{lower}}}$ ▷ Difference ratio
    $v_{adjust} \leftarrow (1 - r_{diff})v_{lower} + r_{diff}v_{upper}$ ▷ Vertical adjustment distance
    $U_{adjusted} \leftarrow U - v_{adjust}$ ▷ Compute adjusted potential
    $A \leftarrow$ area under g curve above $g = 0$ ▷ Computed using trapezoidal rule
    $A_{diff} \leftarrow |1 - A|$ ▷ Deviation from targeted area
    if $A_{diff} > 0$ **then**
      $A_{upper}, v_{upper} \leftarrow A, v_{adjust}$
    else
      $A_{lower}, v_{lower} \leftarrow A, v_{adjust}$
    end if
  end while
  return $U_{adjusted}$
end procedure

Algorithm 4 Pseudocode for the algorithm to compute the overlapping generalised coordination number range between a given generalised coordination number distribution of a given cluster and a given catalytic weighting profile.

Input:

$p : \{p_1, ..., p_n\}$ Generalised coordination numbers distribution of a given cluster

$q : \{q_1, ..., q_n\}$ Catalytic weighting profile

$G : \{g_1, ..., g_n\}$ Generalised coordination numbers in the range of interest

Output:

$O : \{o_1, ..., o_n\}$ Generalised coordination numbers within the overlapping range between p and q

procedure CALCOVERLAP(p, q, G)
    $p_{bulk} \leftarrow p$ ▷ Obtain values within full width at half maximum
    $p_{bulk}[p < \frac{\max(p)}{2}] \leftarrow 0.0$
    $O \leftarrow \emptyset$
    for $g \in G$ **do**
        if $(p_{bulk_i} \neg 0.0)$ & $(q_i \neg 0.0)$ **then**
            $O \leftarrow O \cup g$
        end if
    end for
    return O
end procedure

S7 Correlation matrices

Figures S6 to S18 show the correlation matrices of features for all nanoparticle data sets.

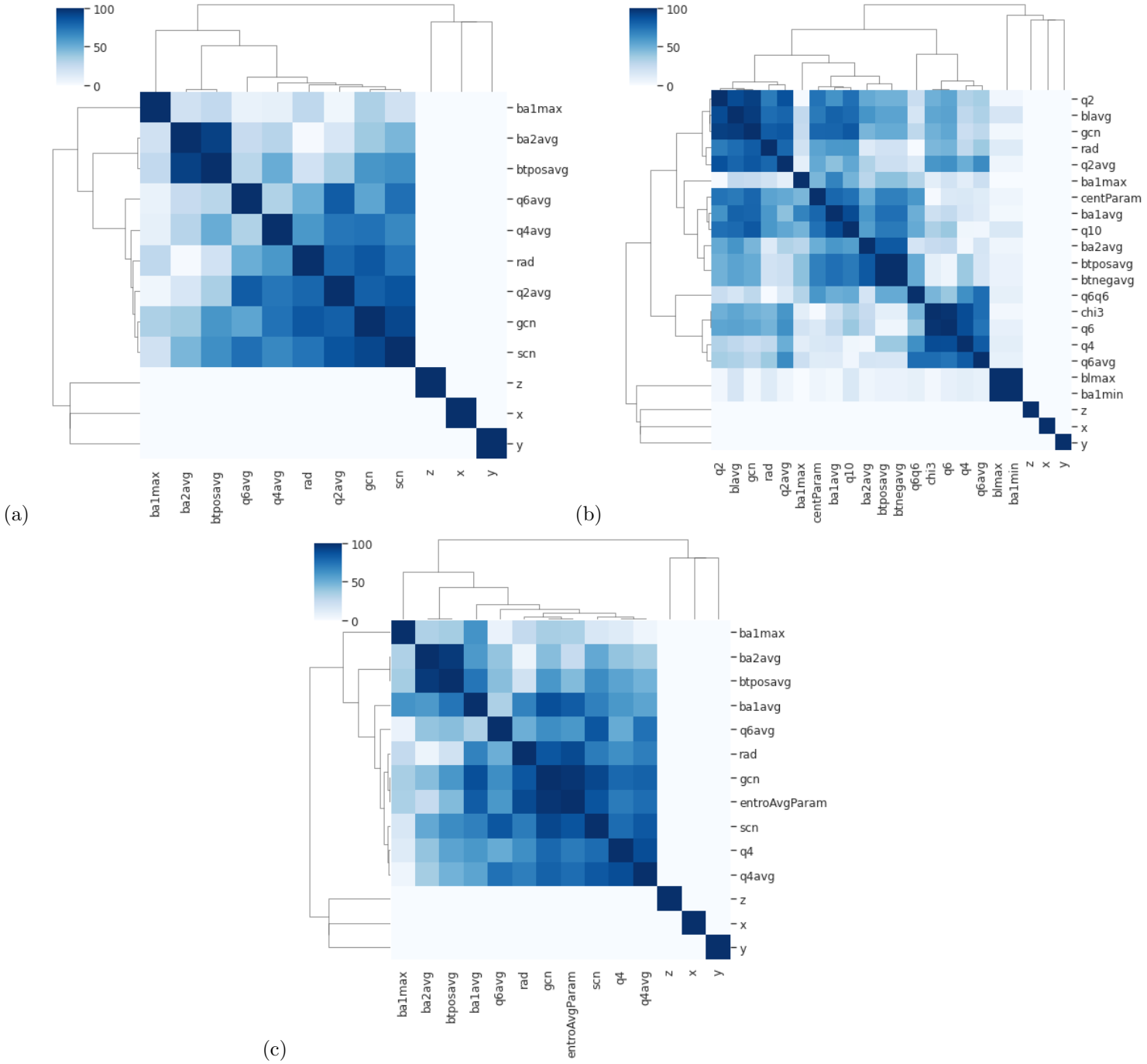


Figure S6: Correlation matrices of features from the data sets of (a) cuboctahedron, (b) icosahedron, and (c) truncated octahedron nanoparticles simulated at 0 K.

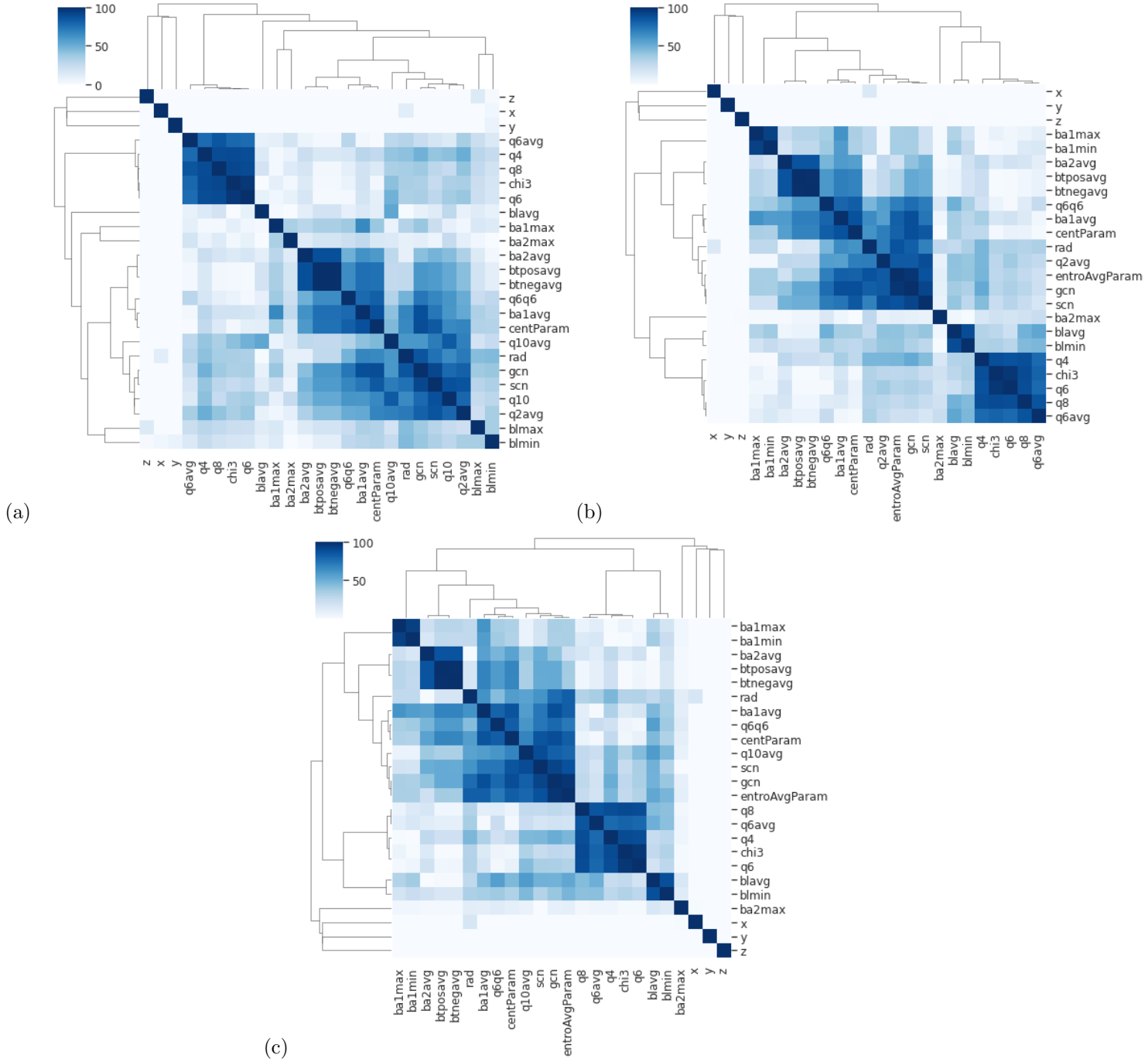


Figure S7: Correlation matrices of features from the data sets of decahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

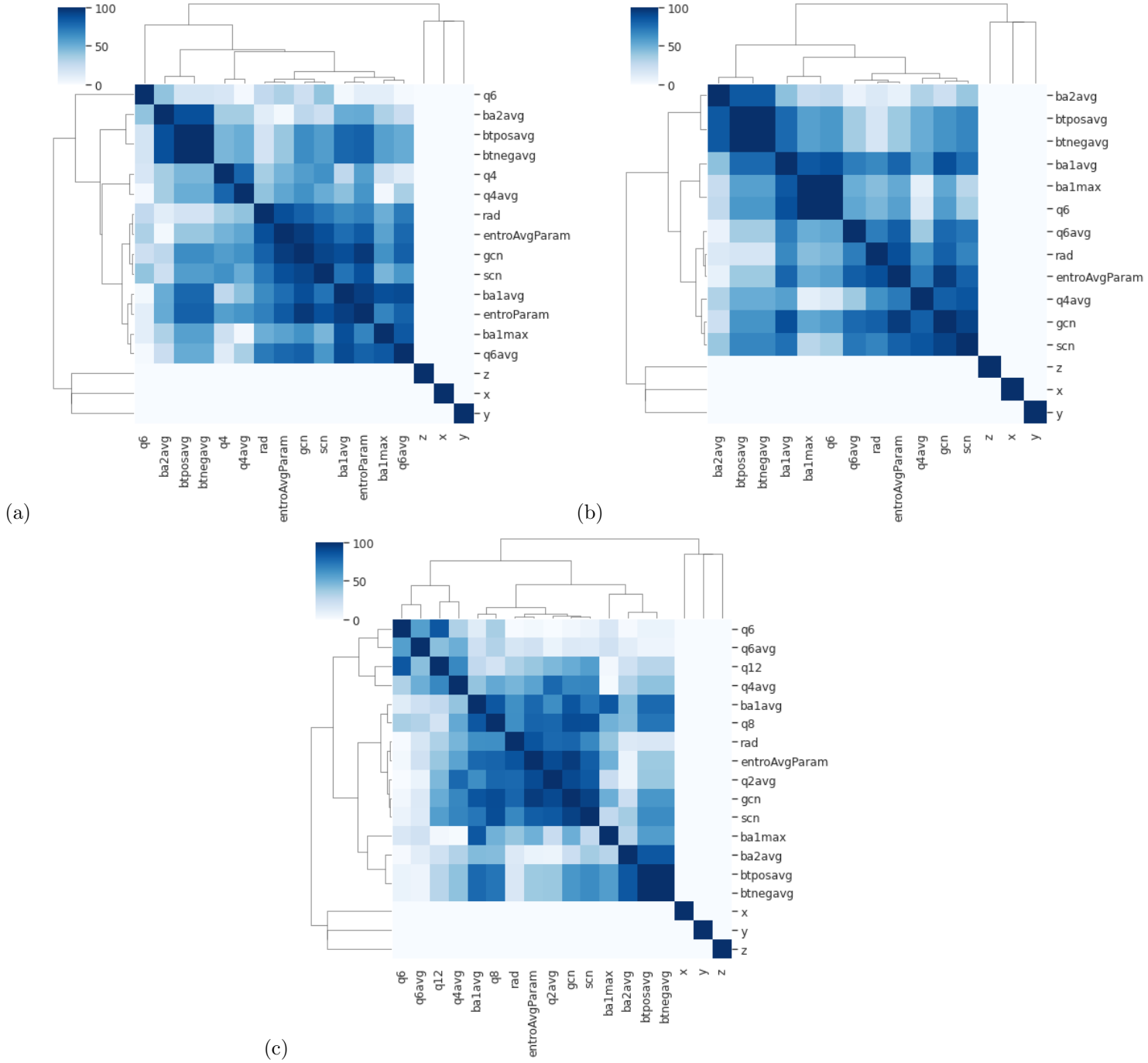


Figure S8: Correlation matrices of features from the data sets of cube nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

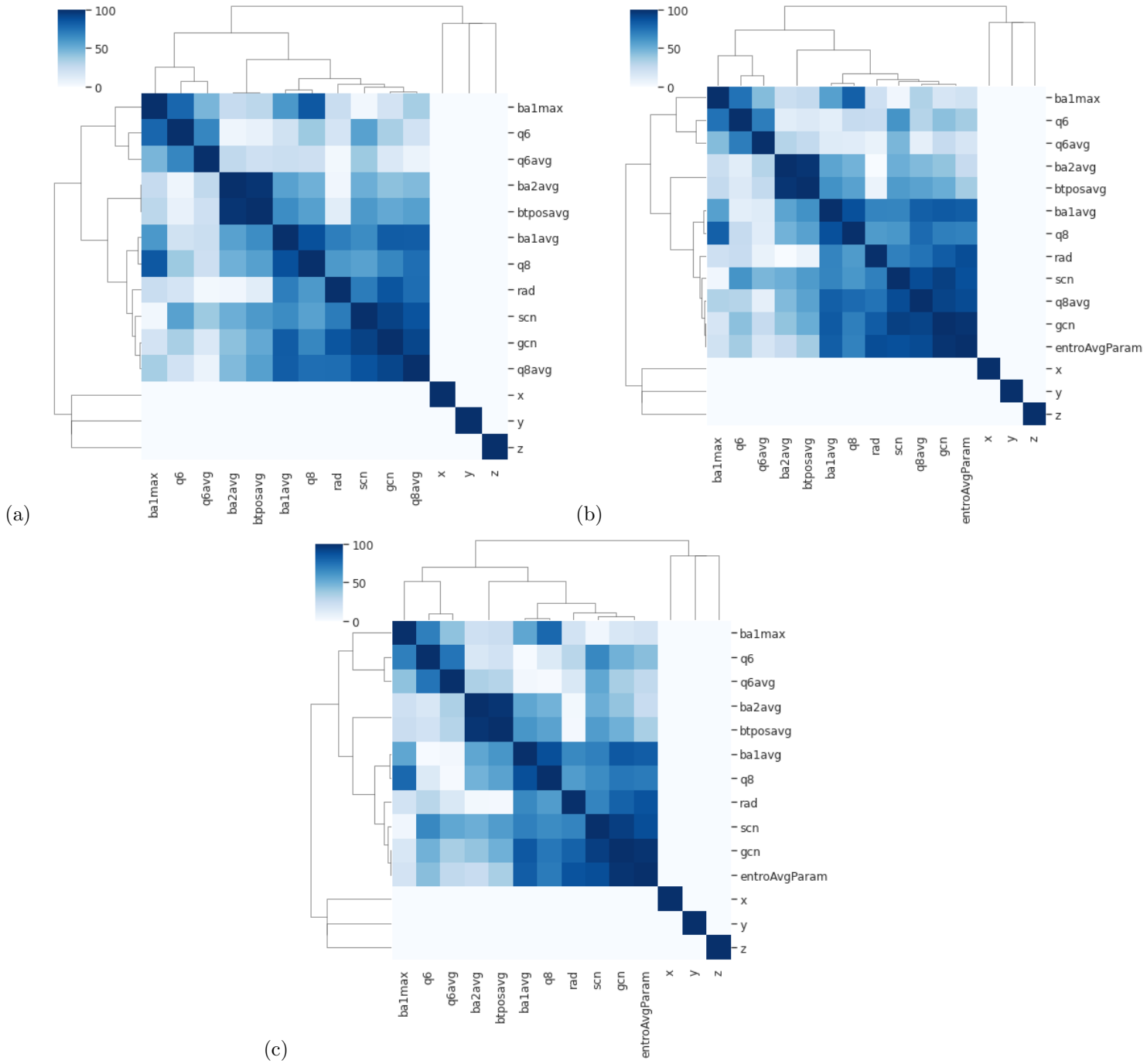


Figure S9: Correlation matrices of features from the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

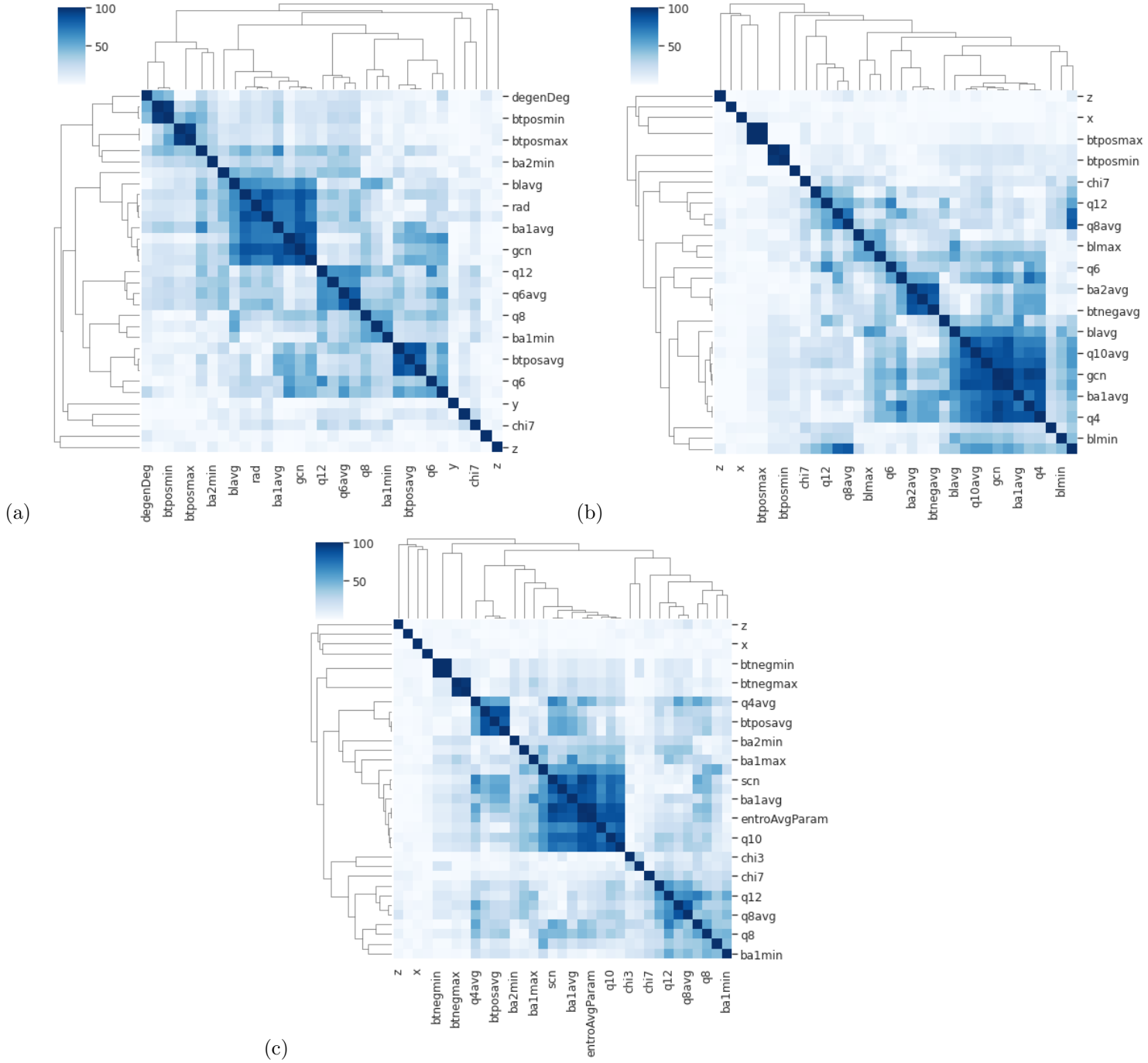


Figure S10: Correlation matrices of features from the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.

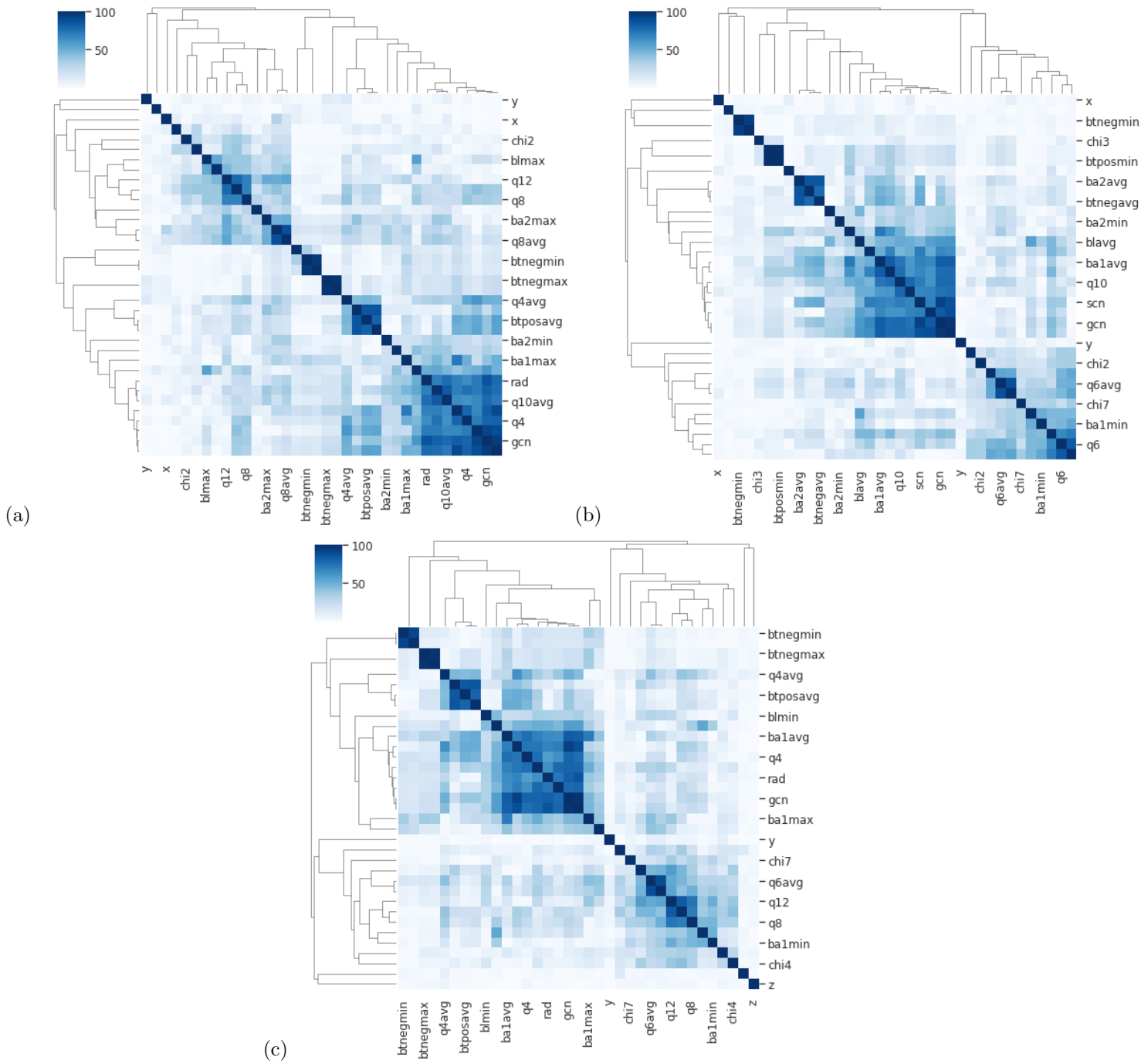


Figure S11: Correlation matrices of features from the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

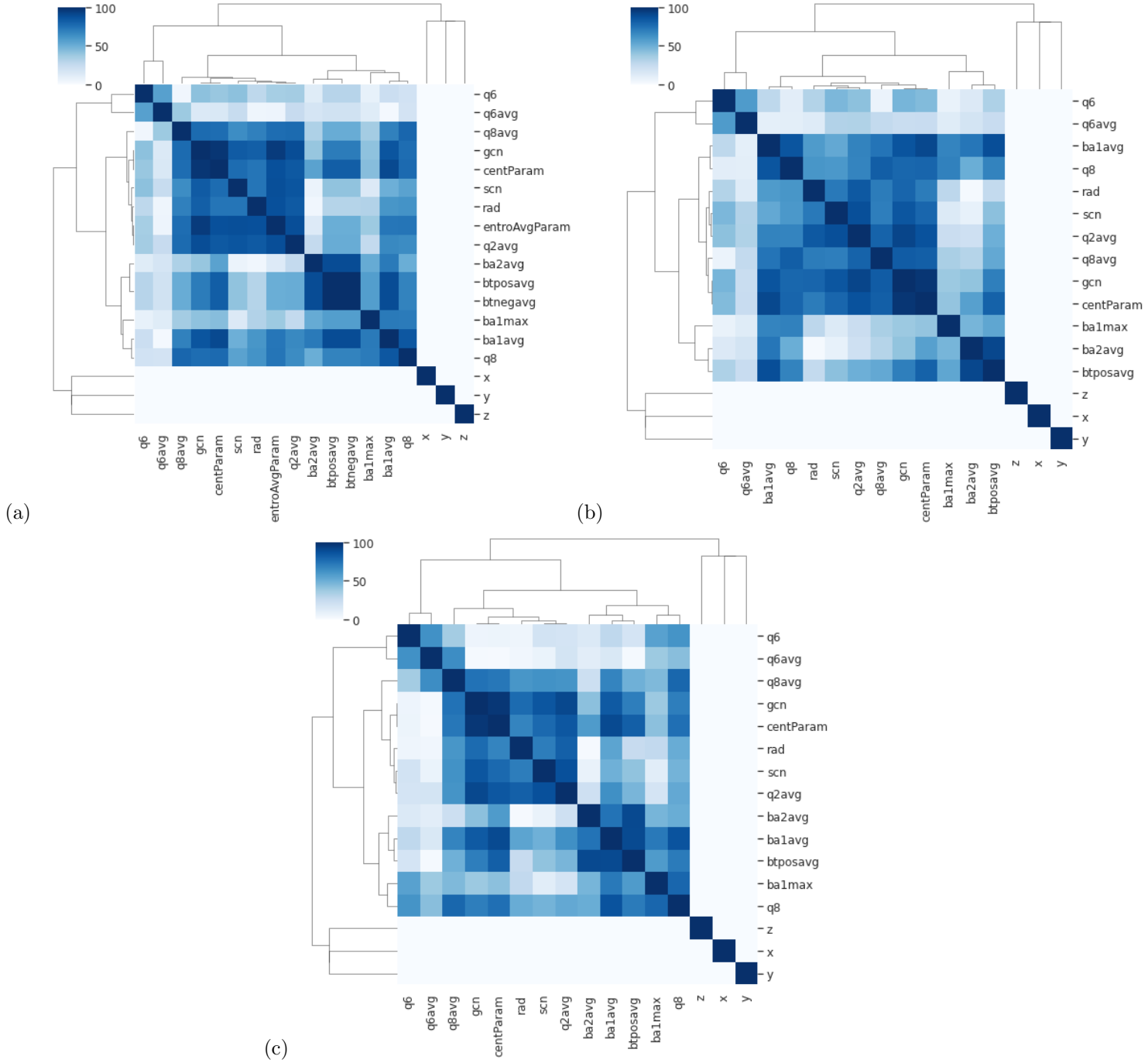


Figure S12: Correlation matrices of features from the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

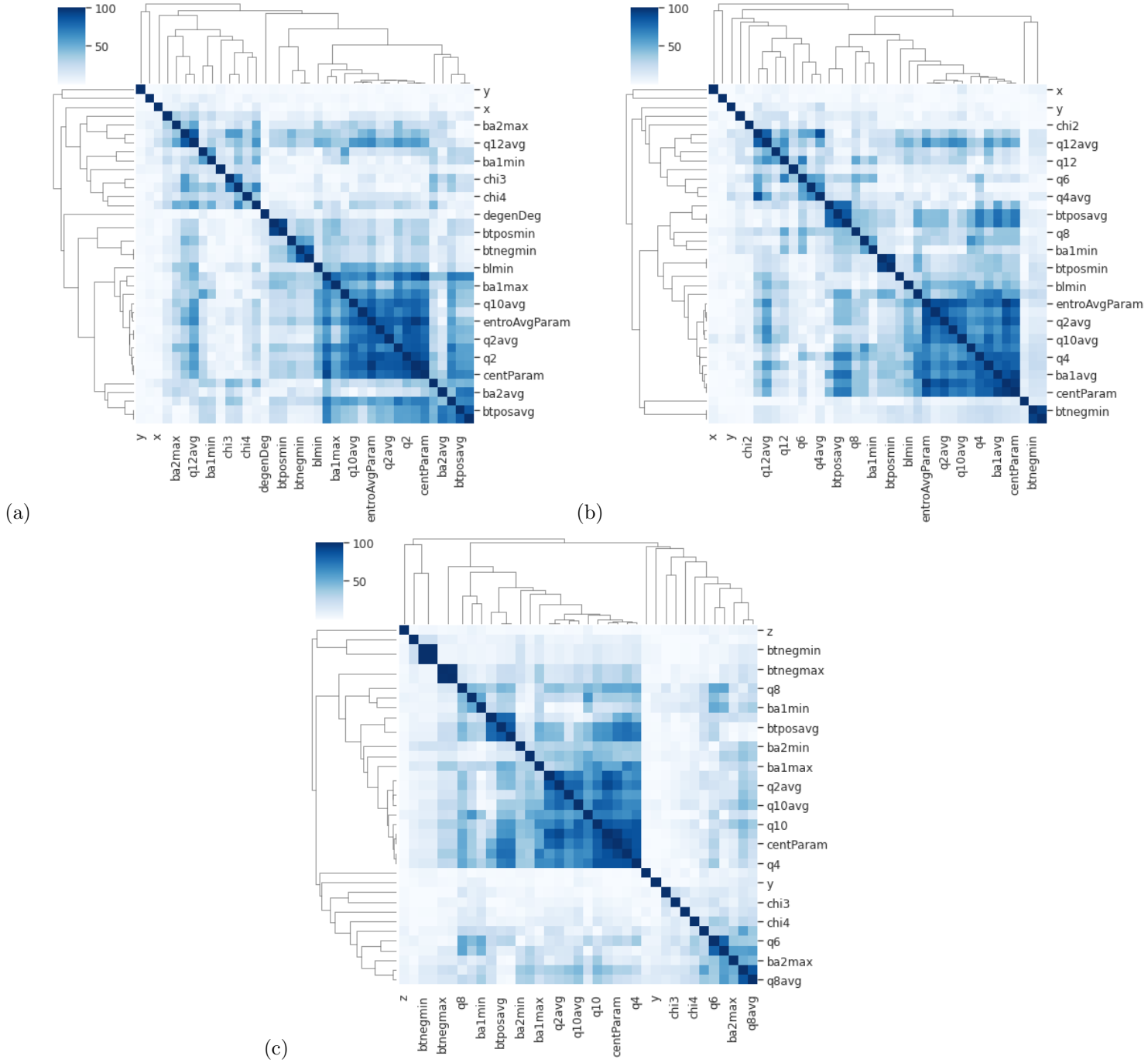


Figure S13: Correlation matrices of features from the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.

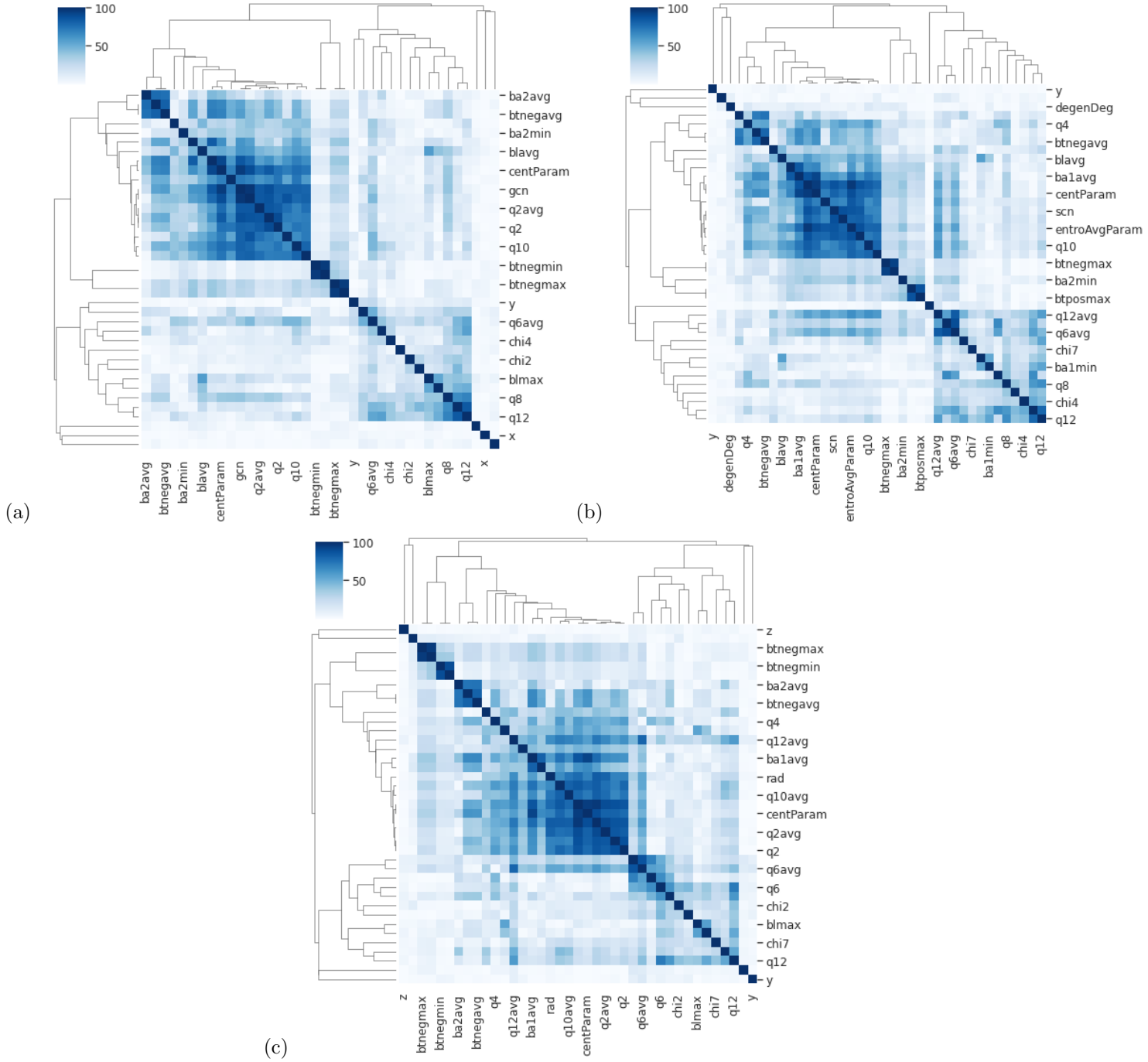


Figure S14: Correlation matrices of features from the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

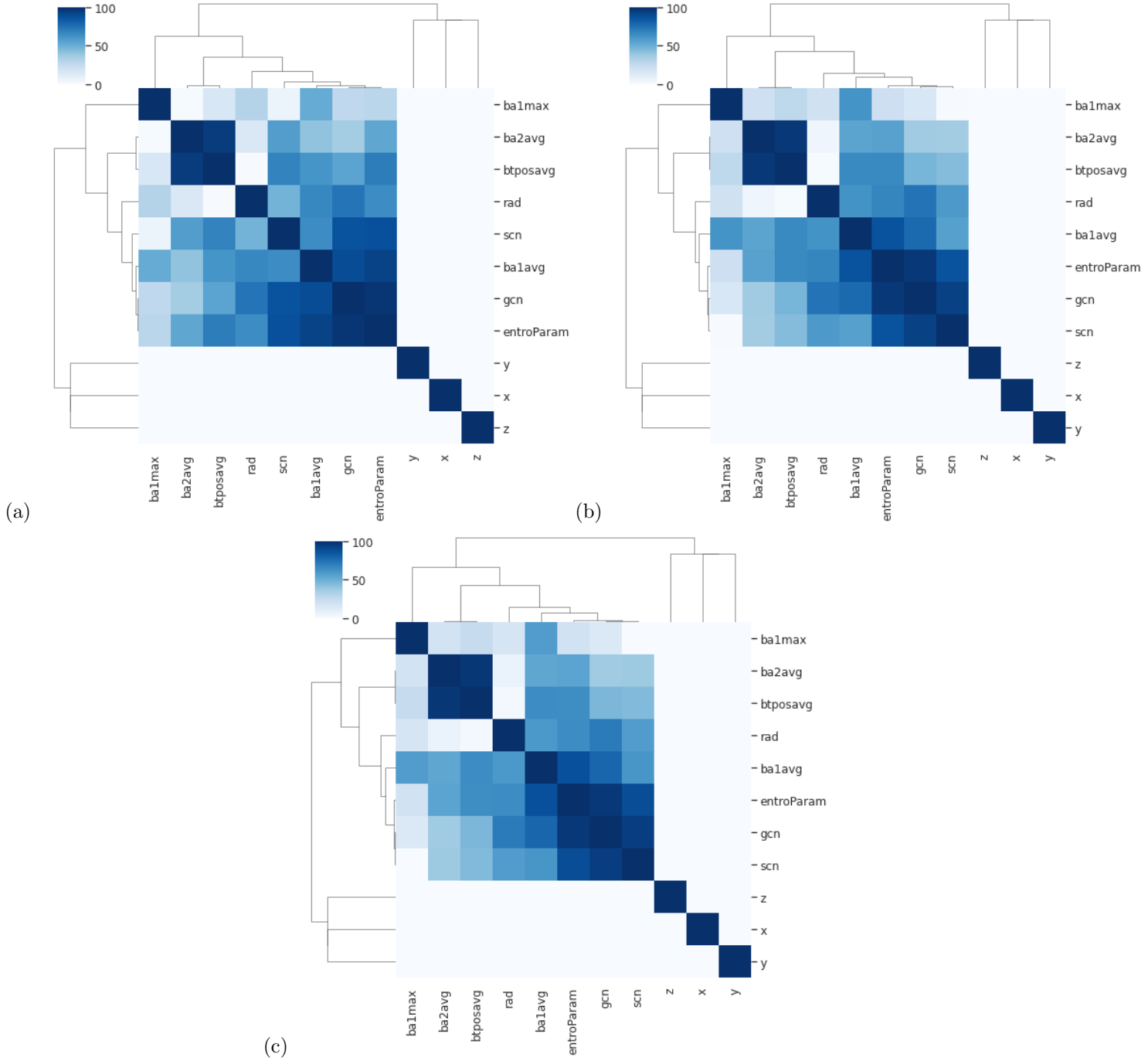


Figure S15: Correlation matrices of features from the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

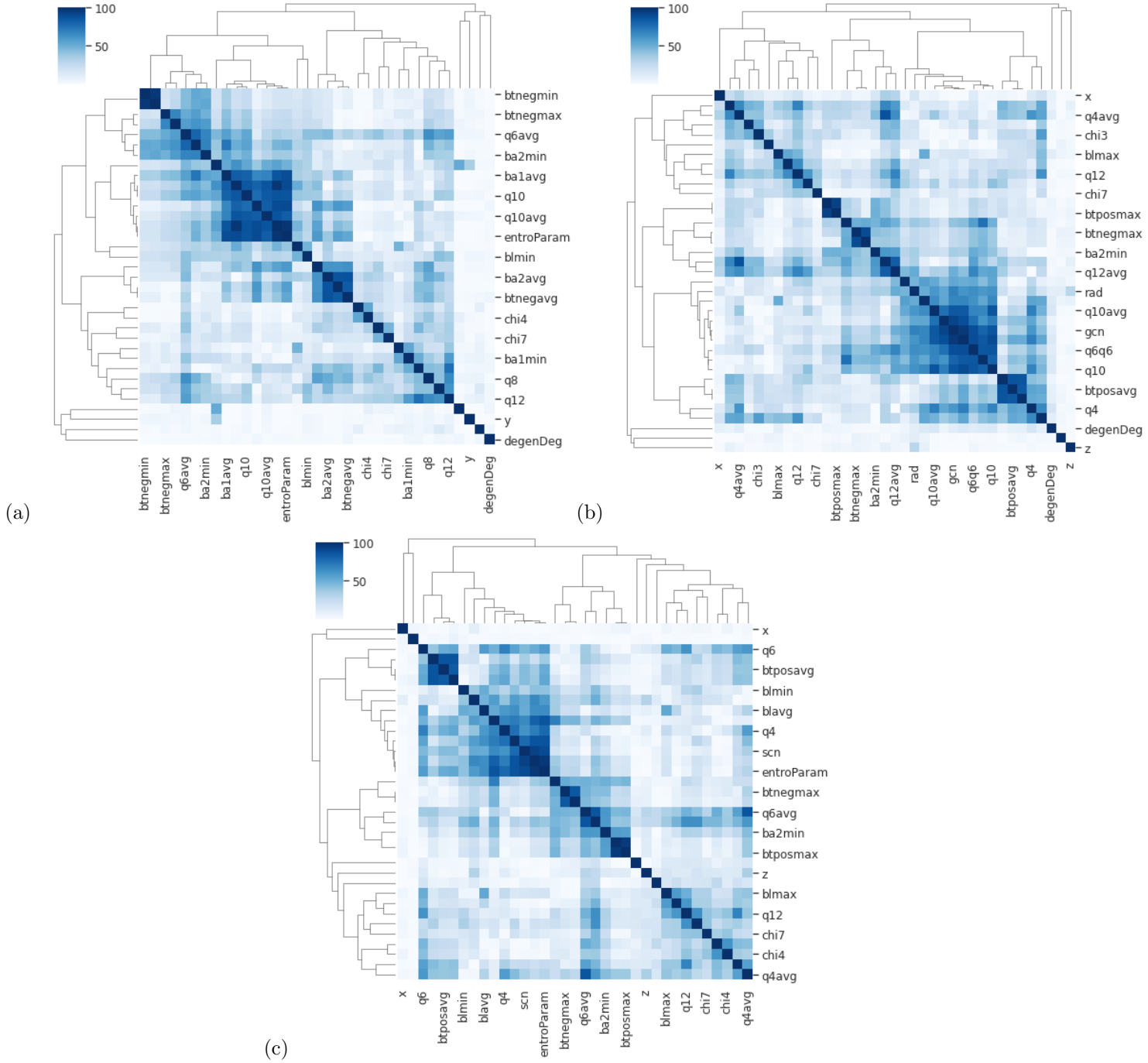


Figure S16: Correlation matrices of features from the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.

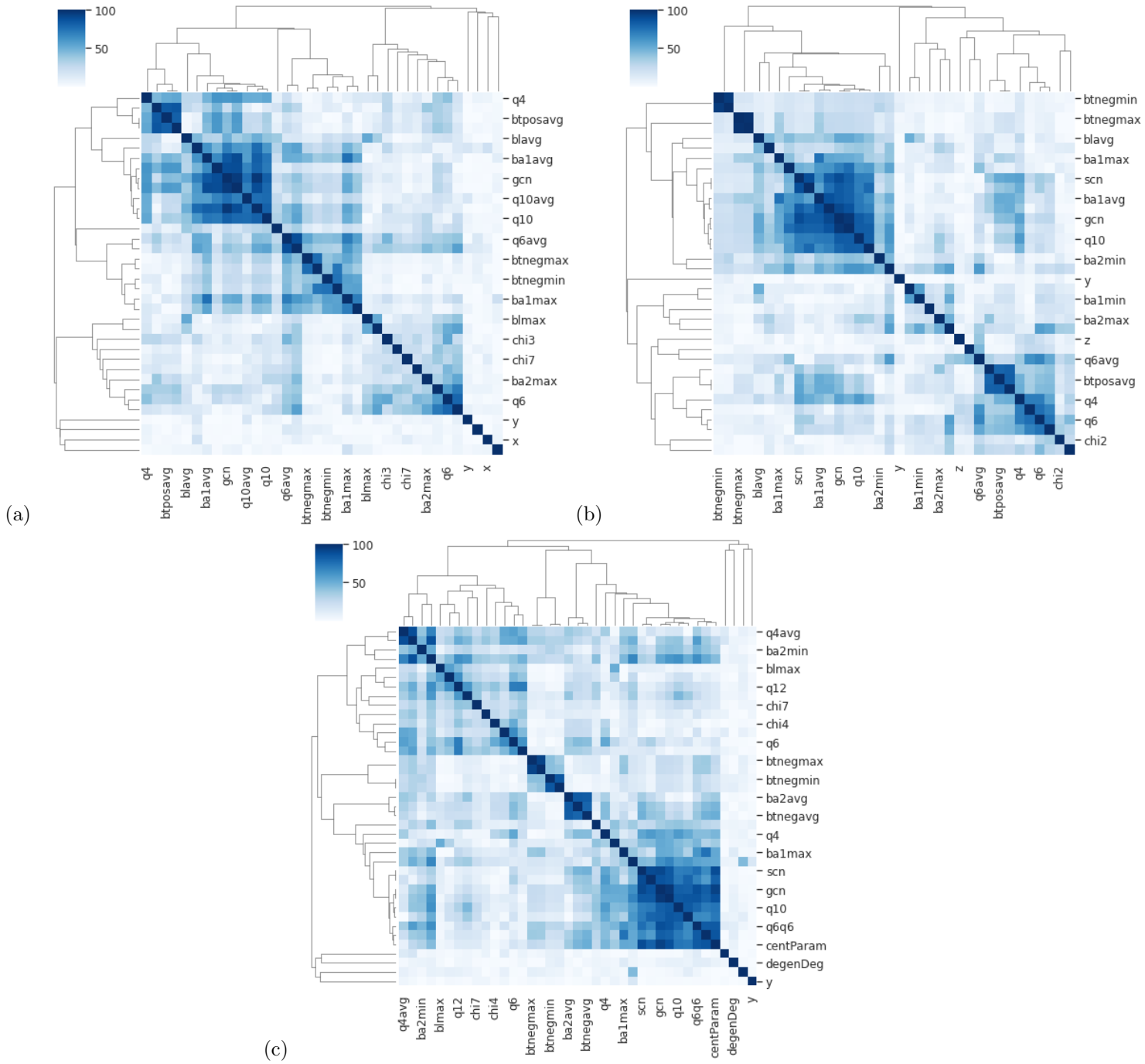


Figure S17: Correlation matrices of features from the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

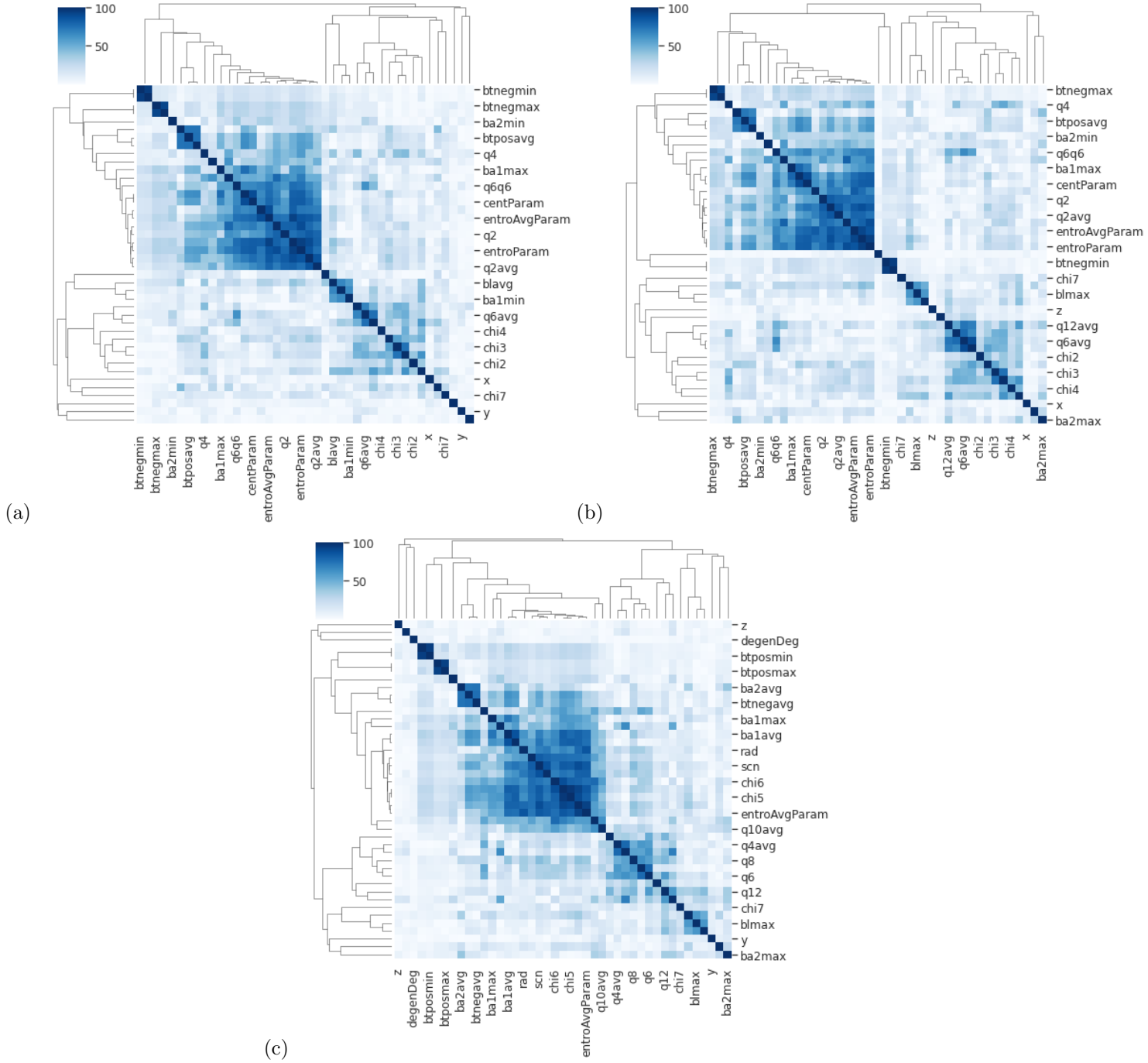


Figure S18: Correlation matrices of features from the data sets of disordered nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 723 K.

S8 Cumulative explained variance plots

Figures S19 to S31 show the cumulative explained variance plots from the principal component analyses of all nanoparticle data sets.

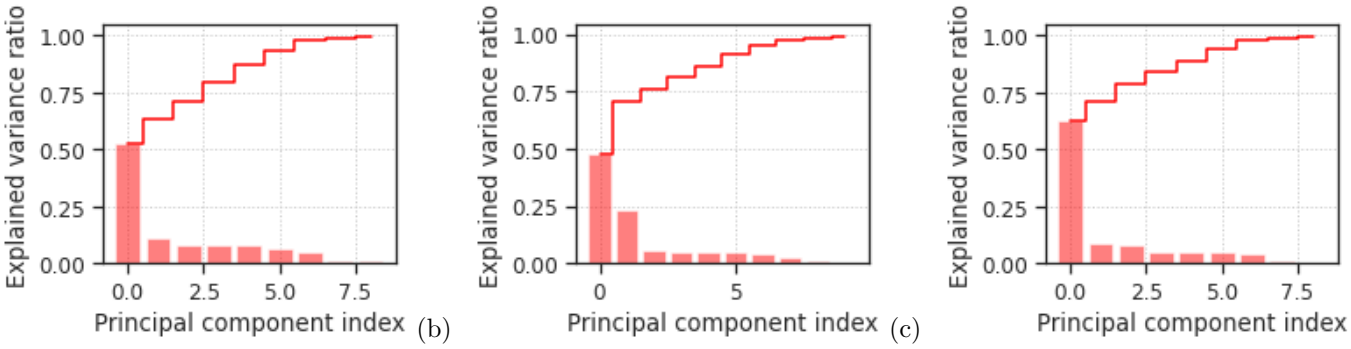


Figure S19: Cumulative explained variance plots from the principal component analyses of the data sets of (a) cuboctahedron, (b) icosahedron, and (c) truncated octahedron nanoparticles simulated at 0 K.

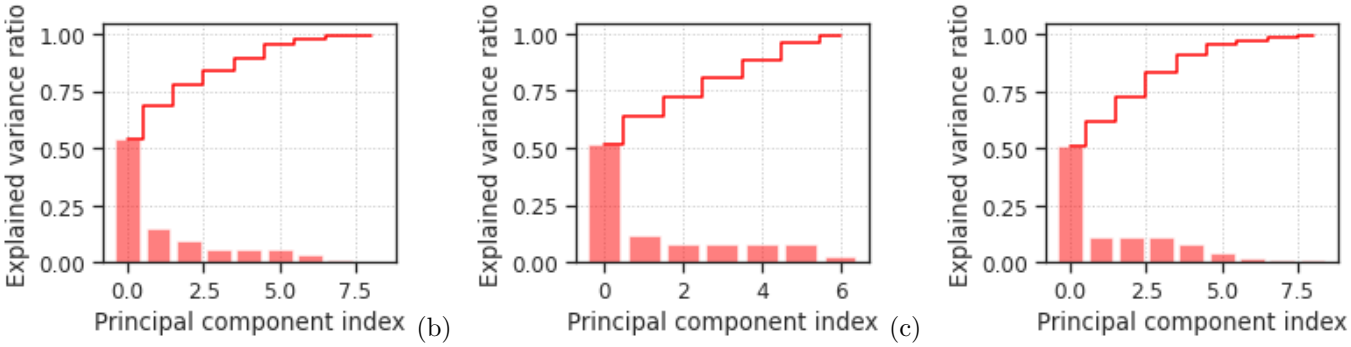


Figure S20: Cumulative explained variance plots from the principal component analyses of the data sets of cube nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

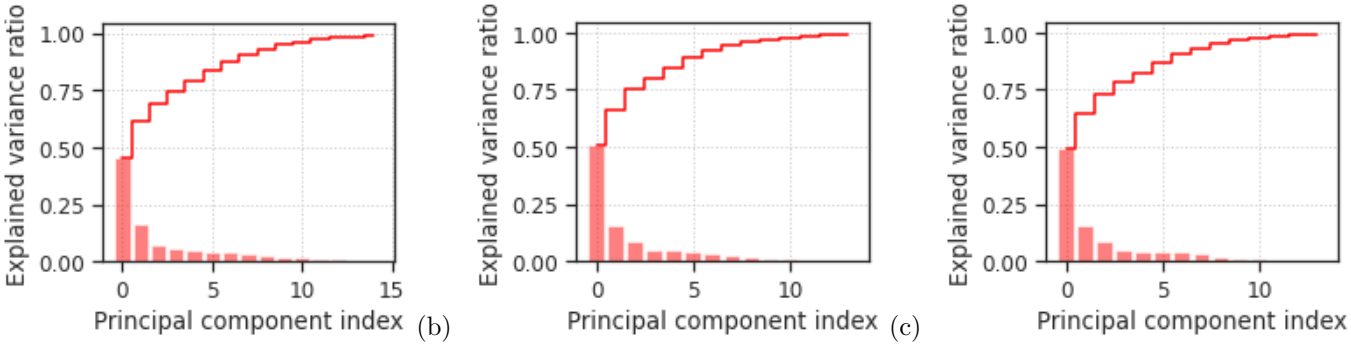
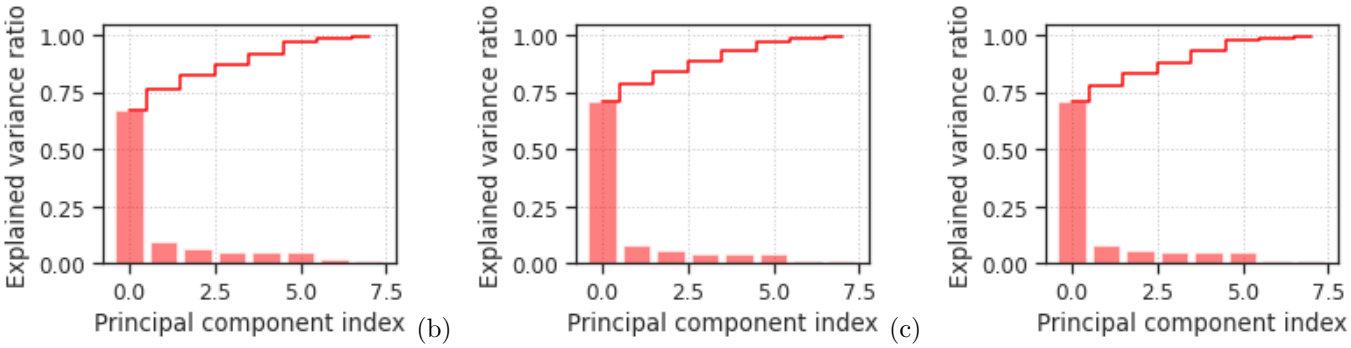
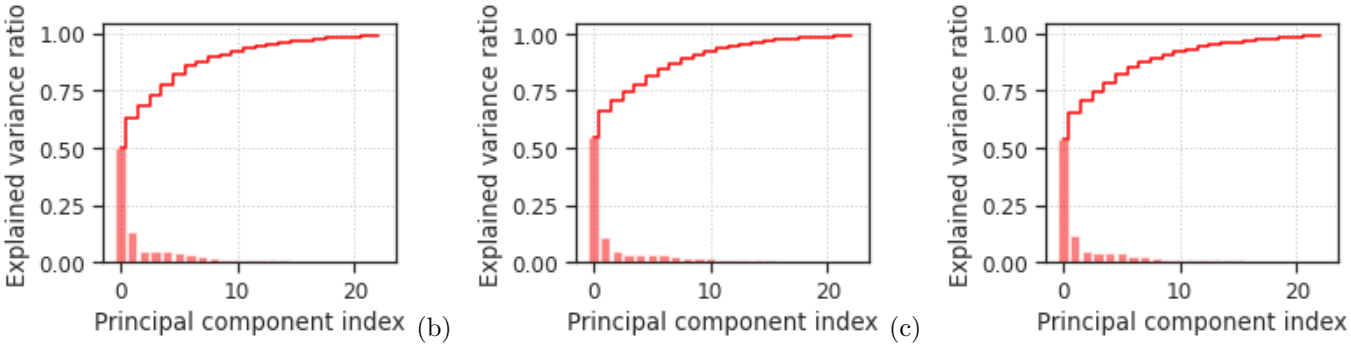


Figure S21: Cumulative explained variance plots from the principal component analyses of the data sets of decahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.



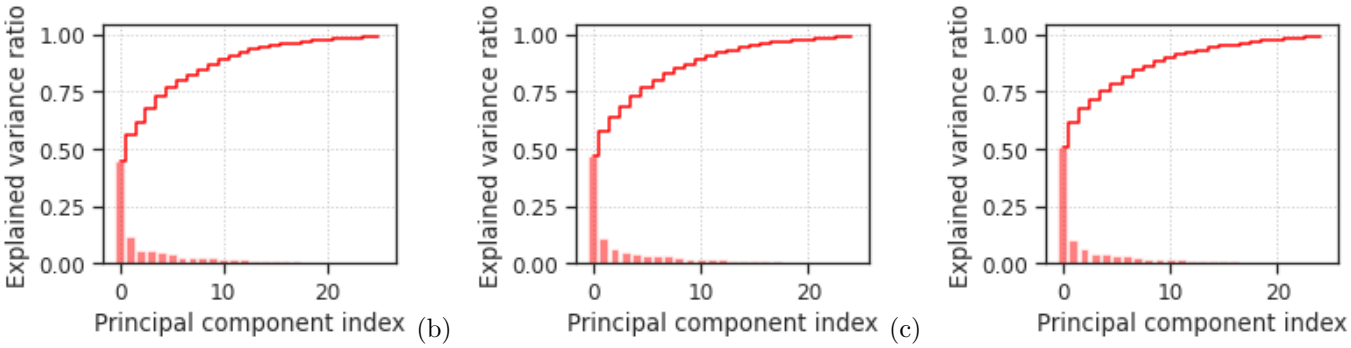
(a) (b) (c)

Figure S22: Cumulative explained variance plots from the principal component analyses of the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.



(a) (b) (c)

Figure S23: Cumulative explained variance plots from the principal component analyses of the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.



(a) (b) (c)

Figure S24: Cumulative explained variance plots from the principal component analyses of the data sets of octahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

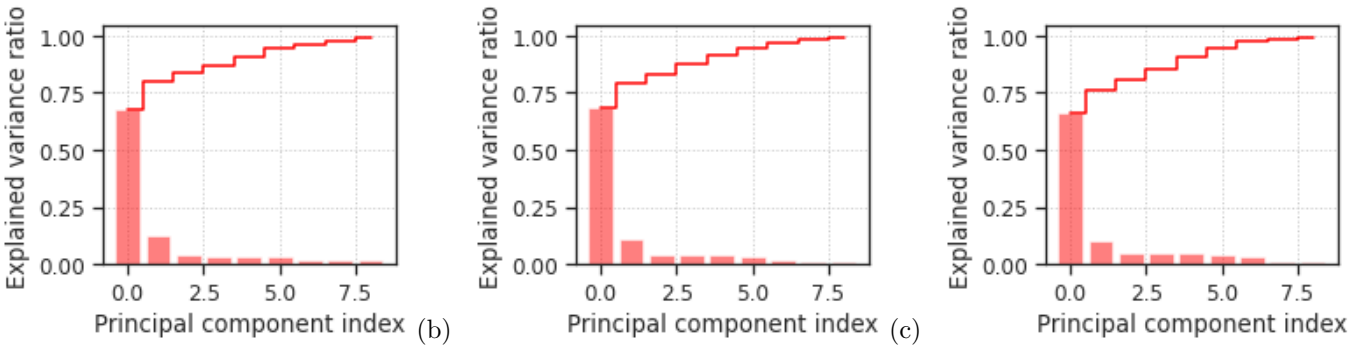


Figure S25: Cumulative explained variance plots from the principal component analyses of the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

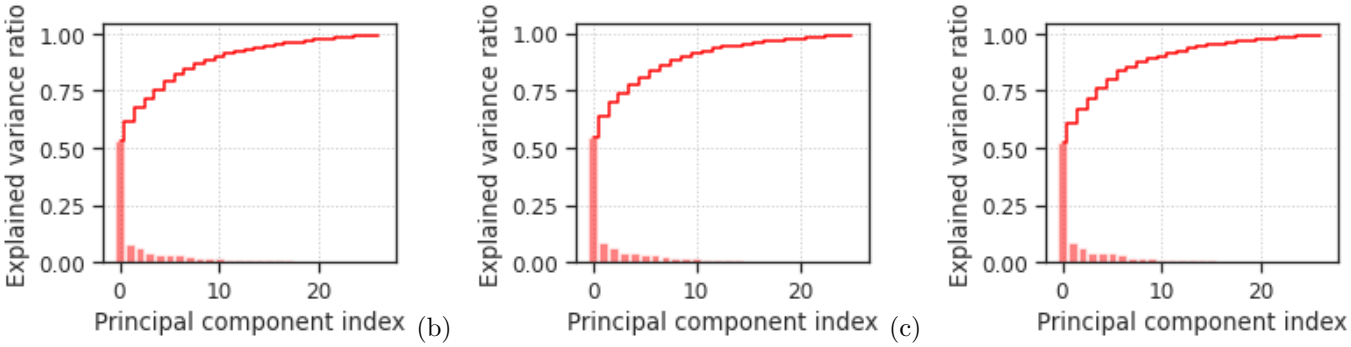


Figure S26: Cumulative explained variance plots from the principal component analyses of the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.

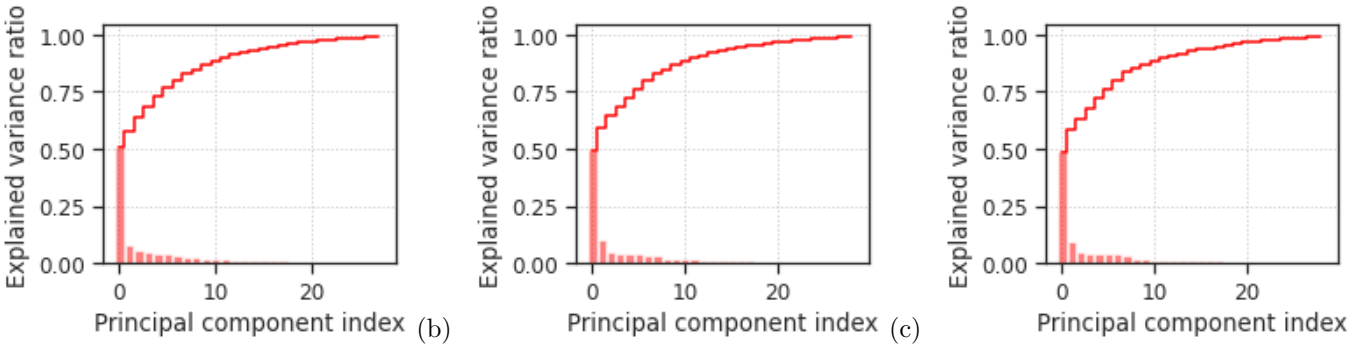


Figure S27: Cumulative explained variance plots from the principal component analyses of the data sets of rhombic dodecahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

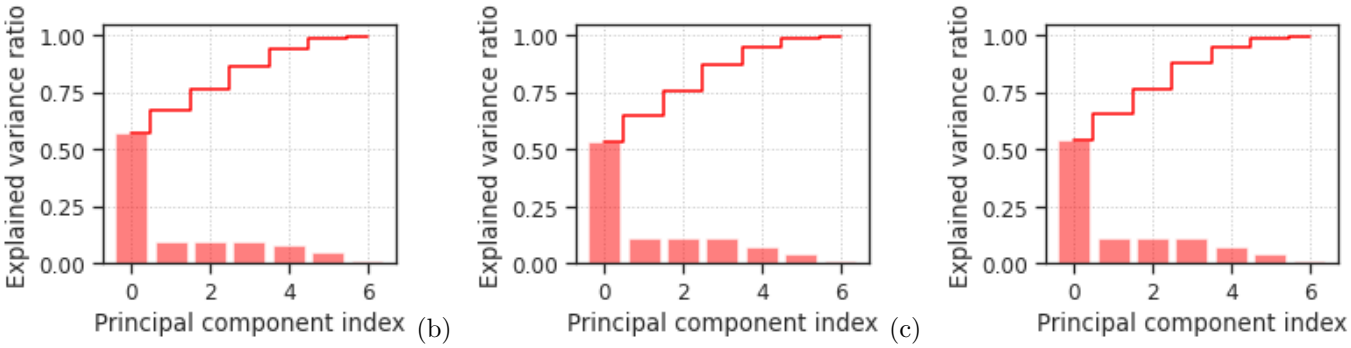


Figure S28: Cumulative explained variance plots from the principal component analyses of the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 0 K.

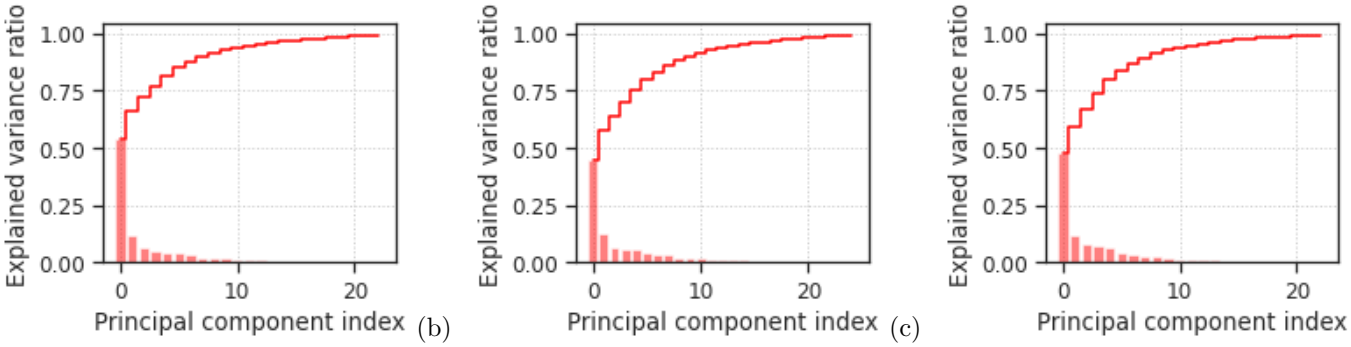


Figure S29: Cumulative explained variance plots from the principal component analyses of the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 323 K.

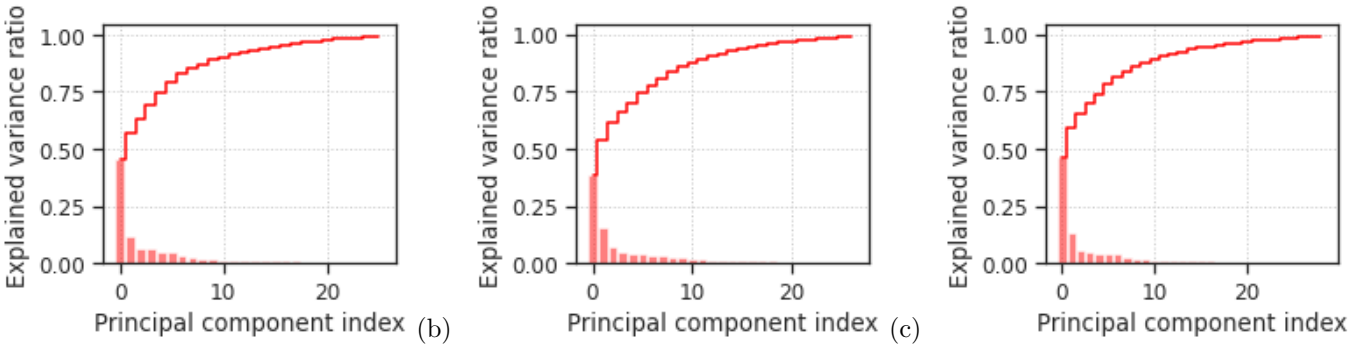


Figure S30: Cumulative explained variance plots from the principal component analyses of the data sets of tetrahedron nanoparticles with (a) small, (b) medium, and (c) large sizes simulated at 523 K.

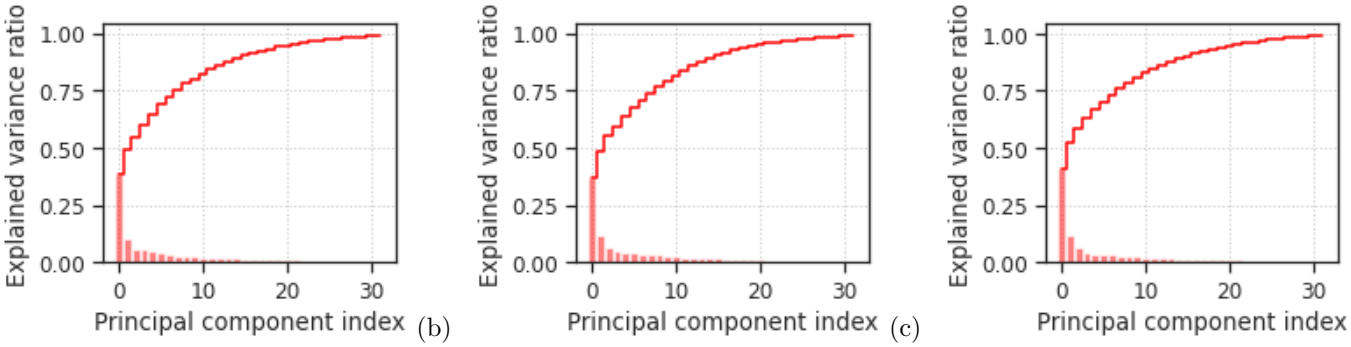


Figure S31: Cumulative explained variance plots from the principal component analyses of the data sets of disordered nanoparticles with (a) small, (b) medium and (c) large sizes simulated at 723 K.

S9 Bulk-surface distinction

The ability to distinguish between the bulk and surface atoms of nanoparticles by ILS is confirmed by comparing Figure S32 with Figure 4(e-h) in the main text. The settings to obtain the central peak are included Section S3.2.

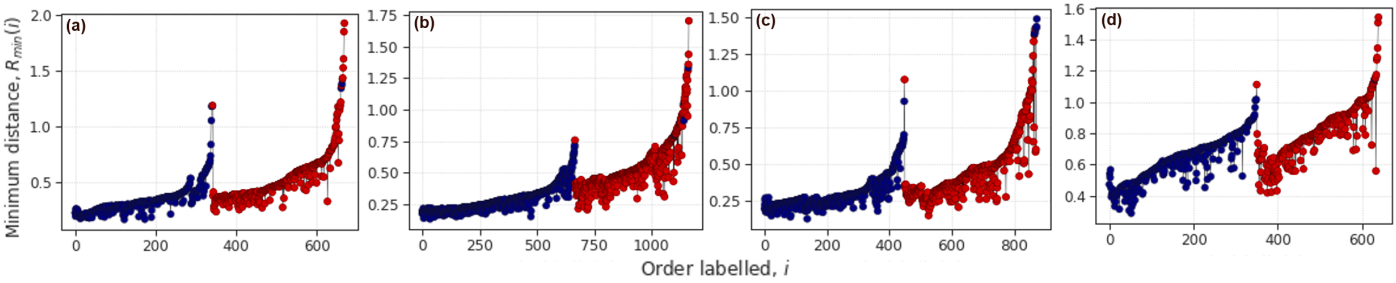


Figure S32: Minimum distance plots for the (a) octahedron, (b) rhombic dodecahedron, and (c) tetrahedron nanoparticles simulated at 523 K, and (d) the smallest disordered nanoparticle simulated at 723 K, coloured by clusters identified.

One exception is the TH nanoparticle simulated at 523 K, where a surface atom identified by the *NCPac*[19] is regarded by the algorithm as closer to the bulk clusters in the current feature set used to describe the atoms, as shown in Figure S33.

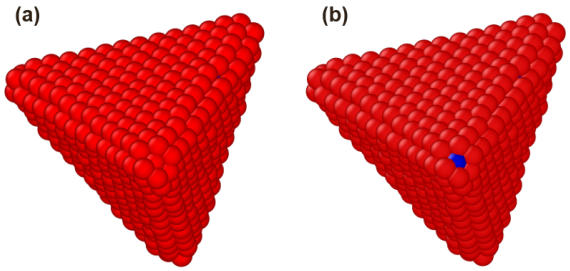


Figure S33: Tetrahedron nanoparticle simulated at 523 K, coloured by (a) whether the atom is identified by *Network Characterisation Package* to be a surface (red) or bulk (blue) atom, and (b) the clusters identified using iterative label spreading.

S10 Subsurface structure distinction

Figure S34 shows that the five-fold twinned axis of the ordered decahedron nanoparticle is successfully distinguished by the clustering algorithm. By tuning the t threshold to a smaller value of 0.1 compared to the default 1.2, more refined details of the subsurface structures can be obtained, as shown in Figure S35 in comparison to Figure 5 in the main text.

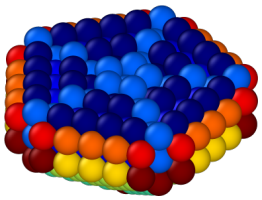


Figure S34: Cross section of the ordered decahedron nanoparticle shown in Figure 5(c) in the main text, slicing across the z-axis or (001) plane. The atoms are coloured by the iterative label spreading clustering results.

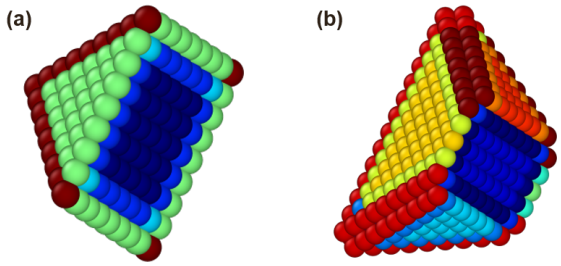


Figure S35: Cross section of the ordered (a) octahedron and (b) tetrahedron nanoparticle shown in Figures 5(e) and 5(g) in the main text, slicing across the x-axis or (100) plane. The atoms are coloured by the iterative label spreading clustering results.

S11 Verification of final clustering results

To confirm that we have obtained the final clustering results, we verified that there is no presence of subclusters by applying the peak identifying algorithm described in Section S3.2 to the R_{min} plots for each cluster of each nanoparticle. To account for the much smaller cluster size, stricter threshold values of k, h, and t are used, being half of the cluster size, 0.3, and 0.01, respectively. As an example, Figure S36 shows the R_{min} plots for each cluster identified on the surface of the disordered RD nanoparticle simulated at 523 K (illustrated in Figure 7(d) in the main text).

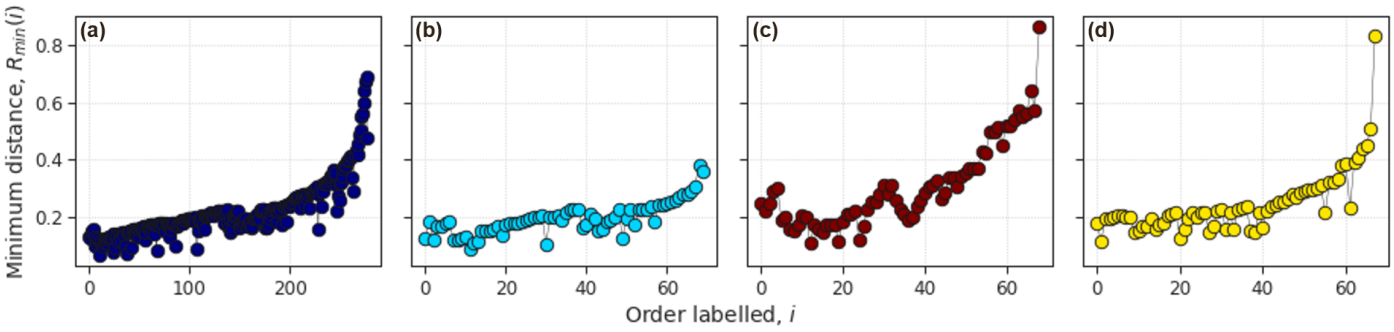


Figure S36: Iterative label spreading minimum distance plots for each cluster identified for the small rhombic dodecahedron nanoparticle simulated at 523 K.

S12 Surface patterns identified using other feature spaces

Figure S37 shows the surface patterns on an octahedron nanoparticle simulated at 523 K (shown in Figure 7(b) in main text) identified by the clustering pipeline with different feature spaces, defined based on the descriptors described in Section S1. Figure S38 shows the ILS minimum distance plots corresponding to 3 different feature spaces, coloured by the identified clusters. It illustrates the impact of the feature space in allowing structural characteristics in disordered nanoparticles to be distinguished. The feature space consisting of the neighbour, order, and Steinhardt descriptors was chosen for the clustering of disordered nanoparticles as the identified surface patterns in Figure S38 best match the expectation of the presence of facets and edges according to the typical structure model.

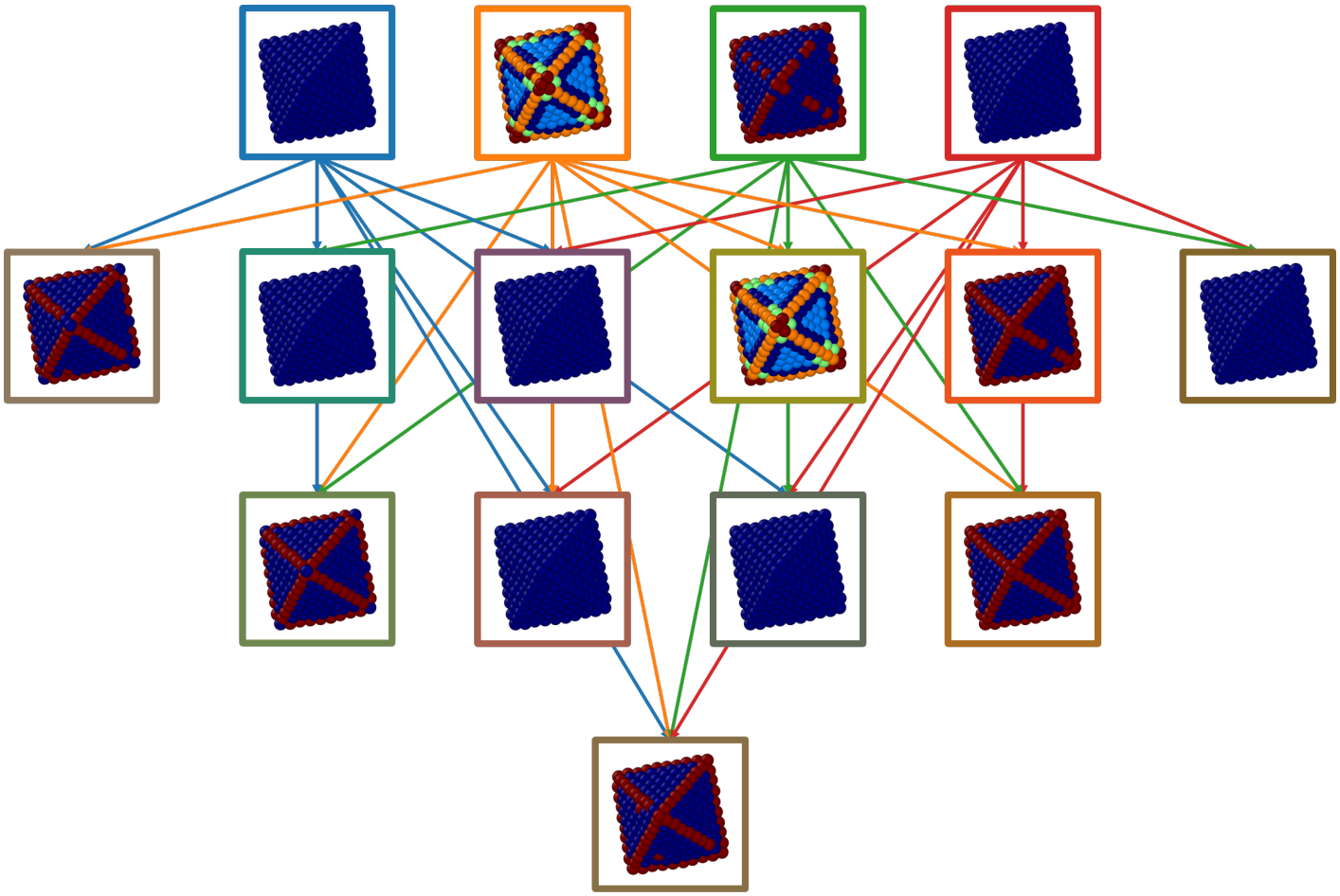


Figure S37: The octahedron nanoparticle simulated at 523 K, coloured by clusters identified from feature spaces consisting of the geometric (blue), neighbour (orange), order (green), and Steinhardt (red) descriptors, and their mixtures. The arrows represent the presence of the descriptors in the feature space.

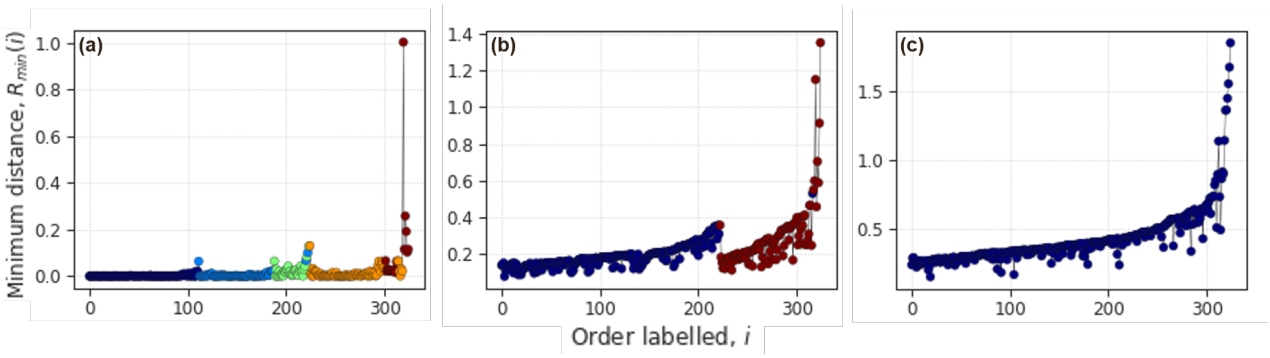


Figure S38: Iterative label spreading minimum distance plots for octahedron nanoparticle simulated at 523 K, coloured by clusters identified from feature spaces consisting of (a) neighbour and order descriptors, (b) neighbour, order, and Steinhardt descriptors, and (c) geometric, order, and Steinhardt descriptors.

S13 Capability of peak identification

Figure S39 shows the peaks and cluster centres identified in the R_{min} plots obtained from the initial ILS run for the surface atoms of an ordered DH and a DIS Pd nanoparticle, showing that the peak-finding algorithm is able to identify the peaks where human eyes might fail (see Figure S39(b)). The cluster centres (the sample with the highest γ score) identified by the algorithm are not always in the middle of the cluster as the computation of γ takes into account a sample's proximity to other high density samples apart from involving the local density of the given sample. This can be understood as the cluster having a diffusive shape, where the members of the cluster lie increasingly further away from each other and the cluster centre.

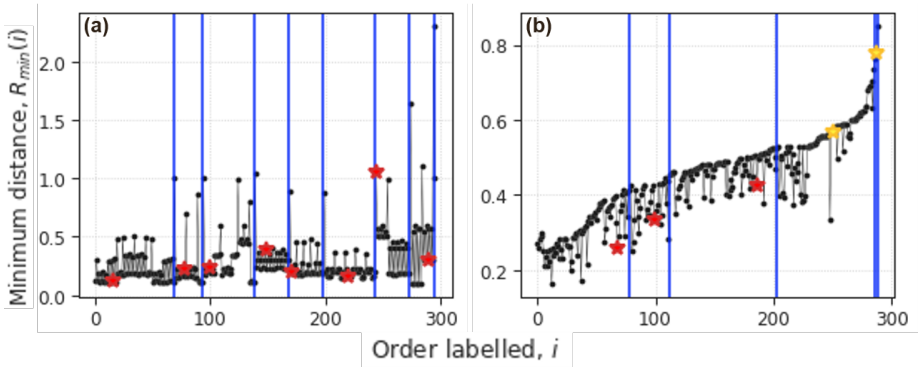
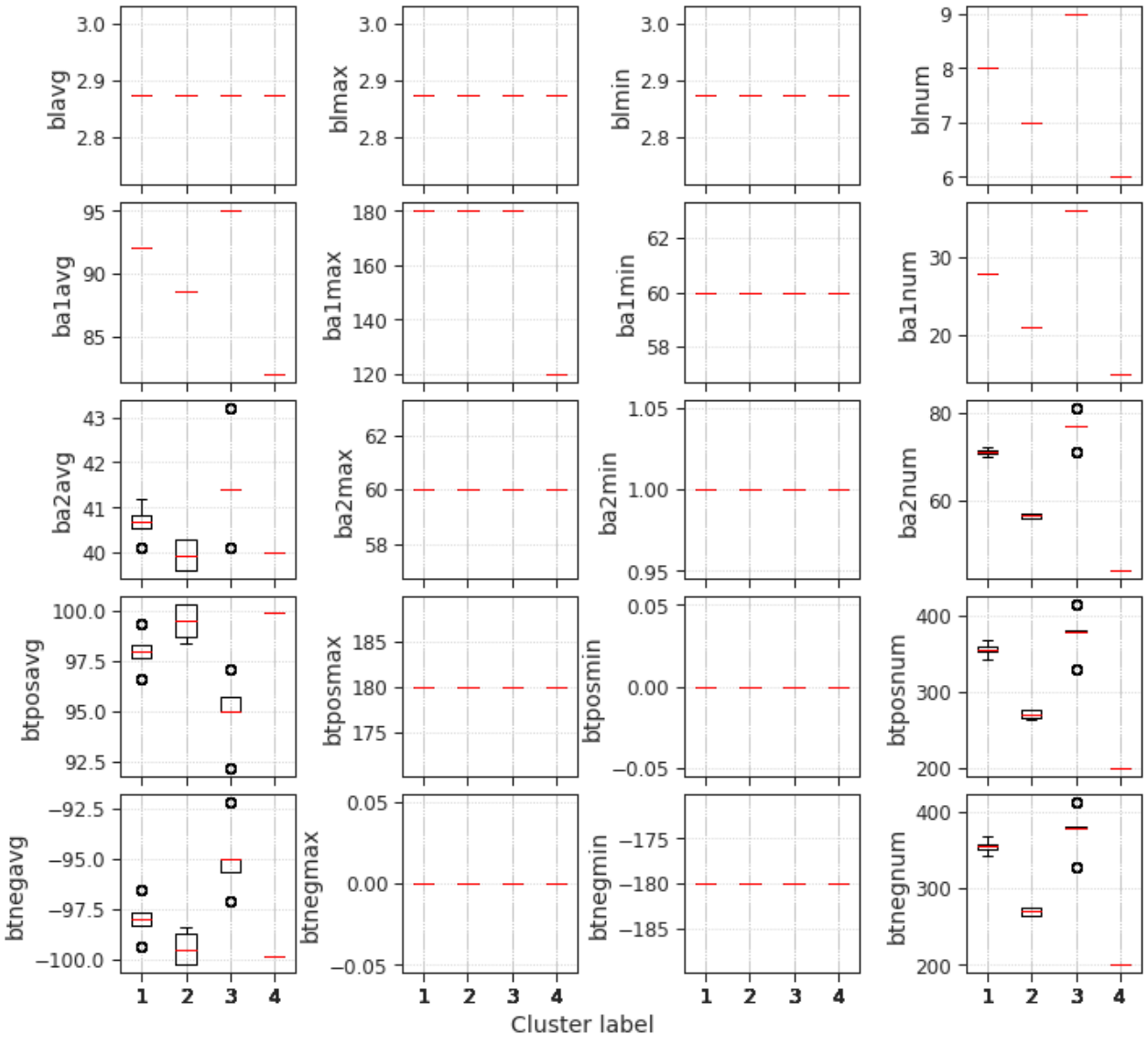


Figure S39: Peaks identified from the iterative label spreading minimum distance plots for a (a) decahedron nanoparticle simulated at 0 K and (b) small-sized disordered nanoparticles simulated at 723 K. The red stars and blue lines correspond to the cluster centres and end points, respectively. The yellow stars indicate that the clusters that they belong to are merged into the closest clusters.

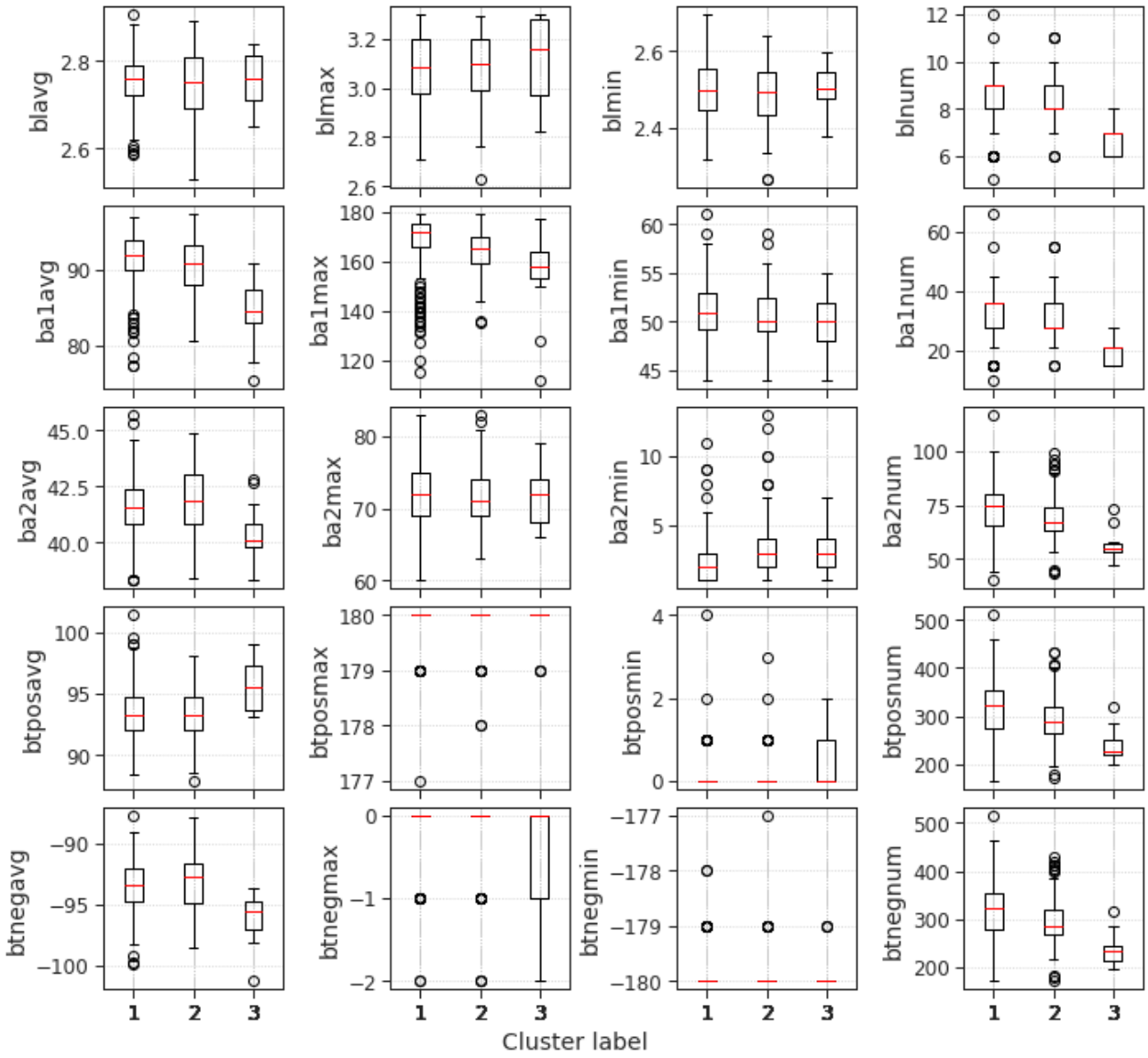
S14 Surface cluster feature profiles

Figures S40 to S44 shows the box plots that profile the features for all surface atom clusters identified for the ordered cuboctahedron nanoparticle simulated at 0 K, and the smallest disordered nanoparticle simulated at 723 K.



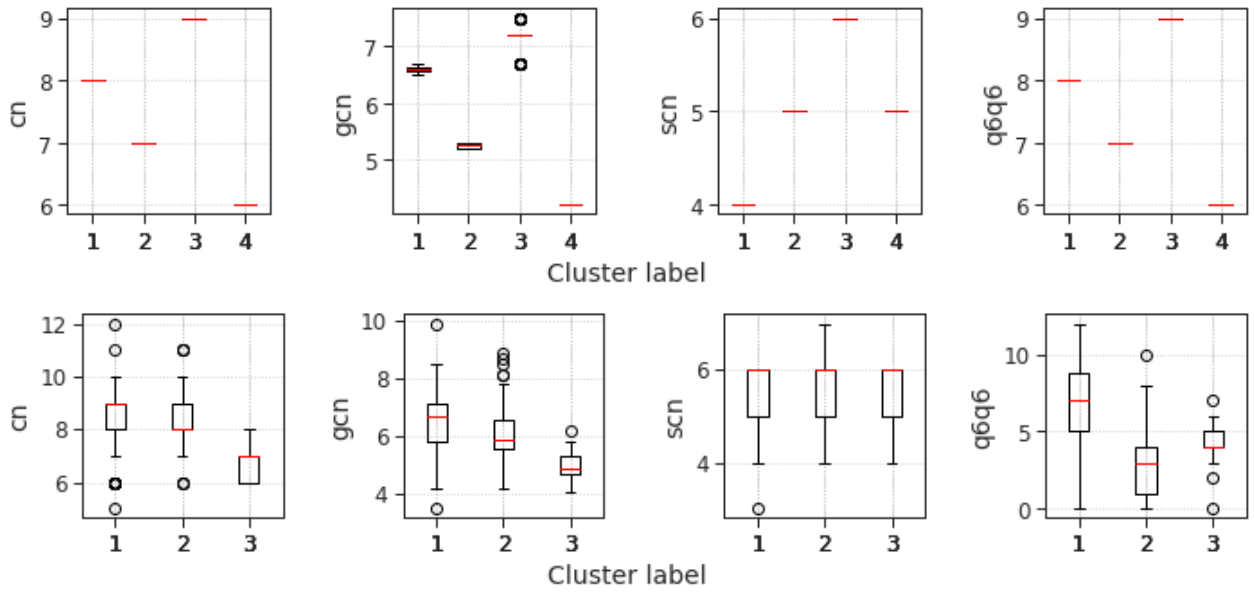
(a)

Figure S40: Box plots of features related to bond geometry for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.



(b)

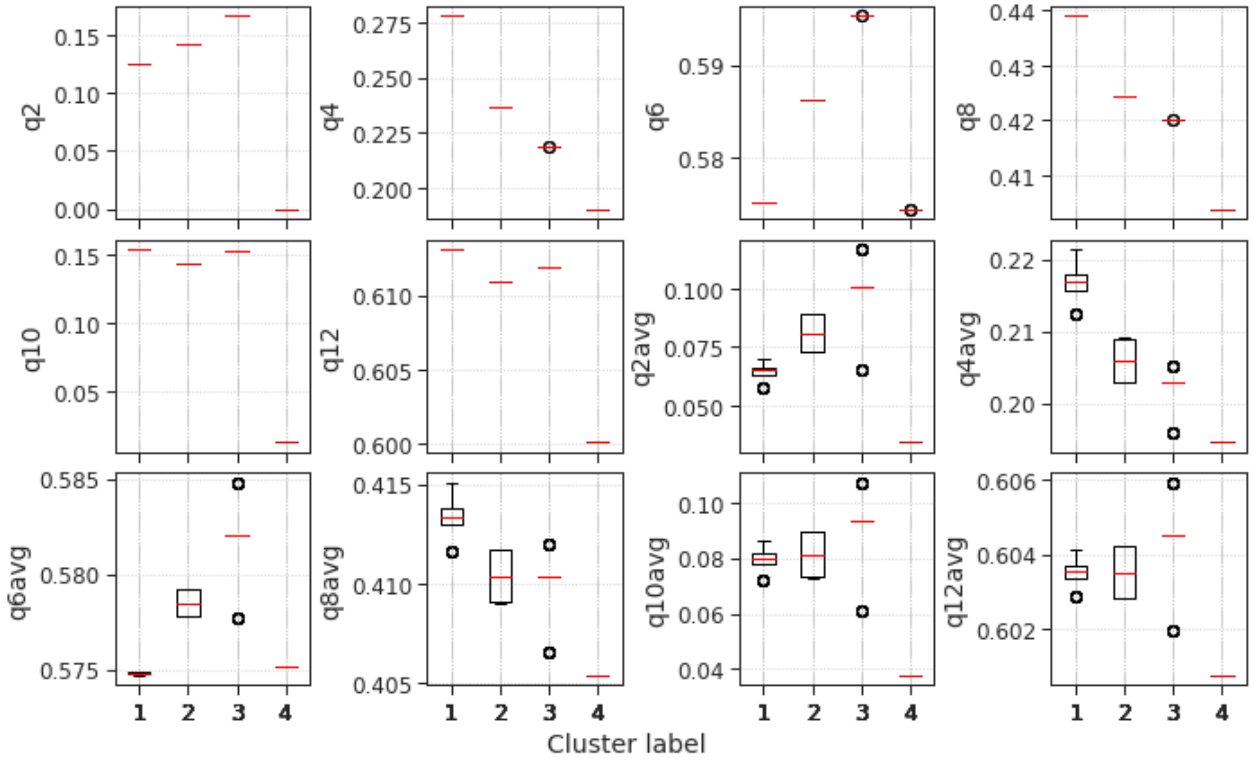
Figure S40: Box plots of features related to bond geometry for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.



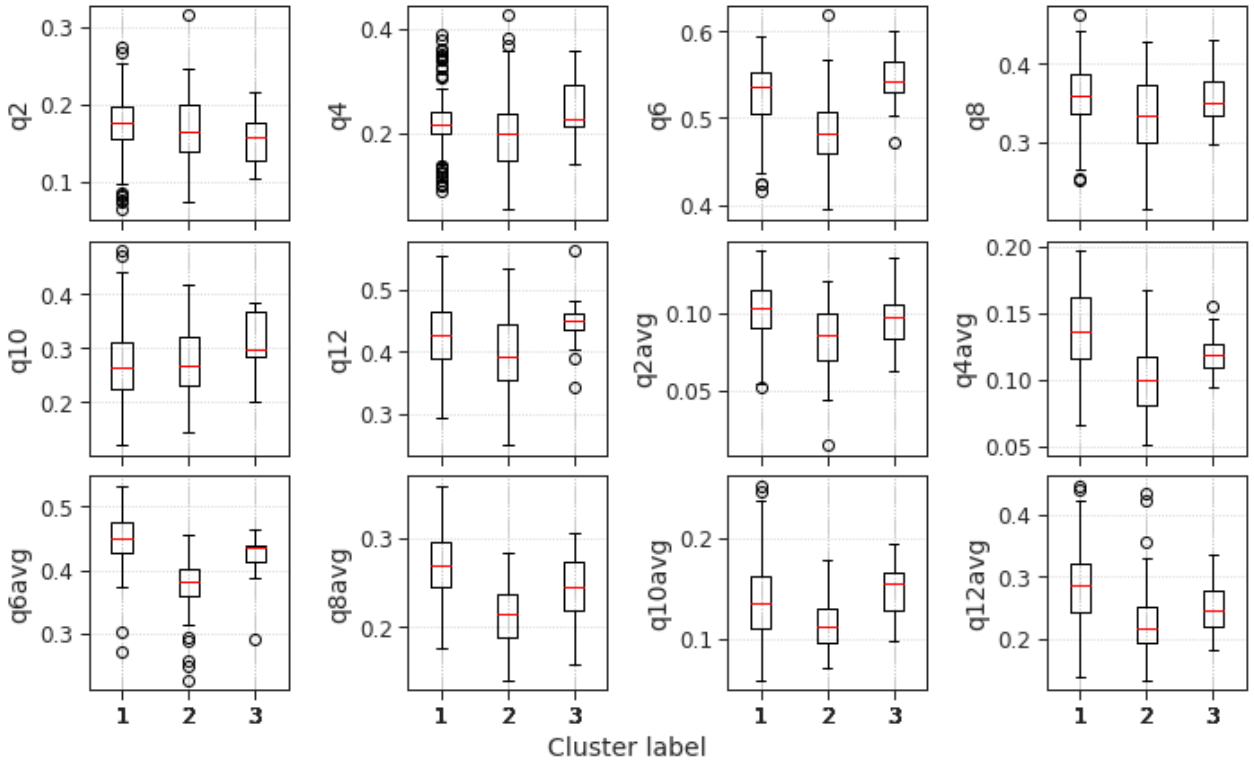
(a)

(b)

Figure S41: Box plots of features related to atom coordination for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.



(a)



(b)

Figure S42: Box plots of features related to Steinhardt's parameters for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.

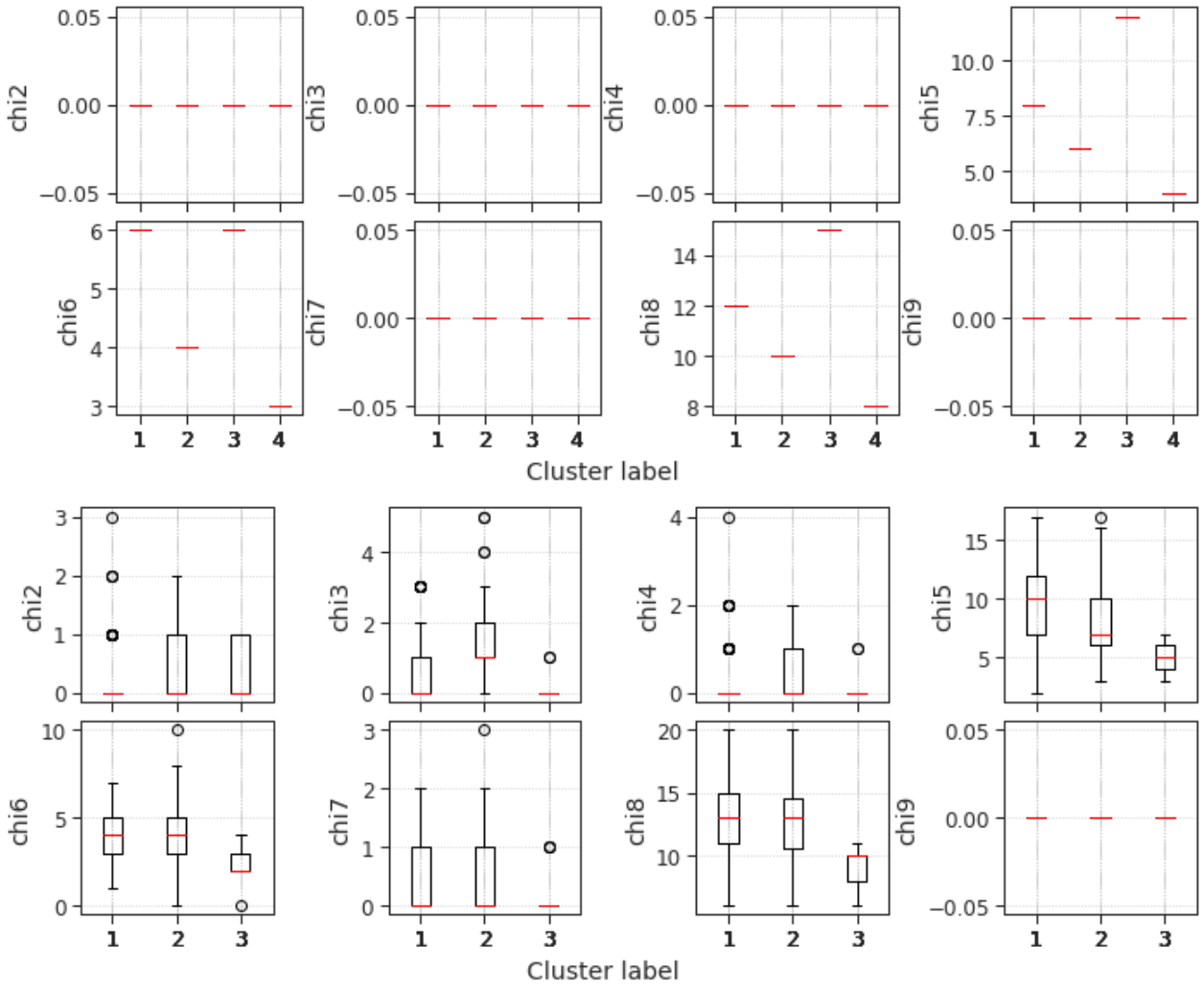


Figure S43: Box plots of features related to χ parameters for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.

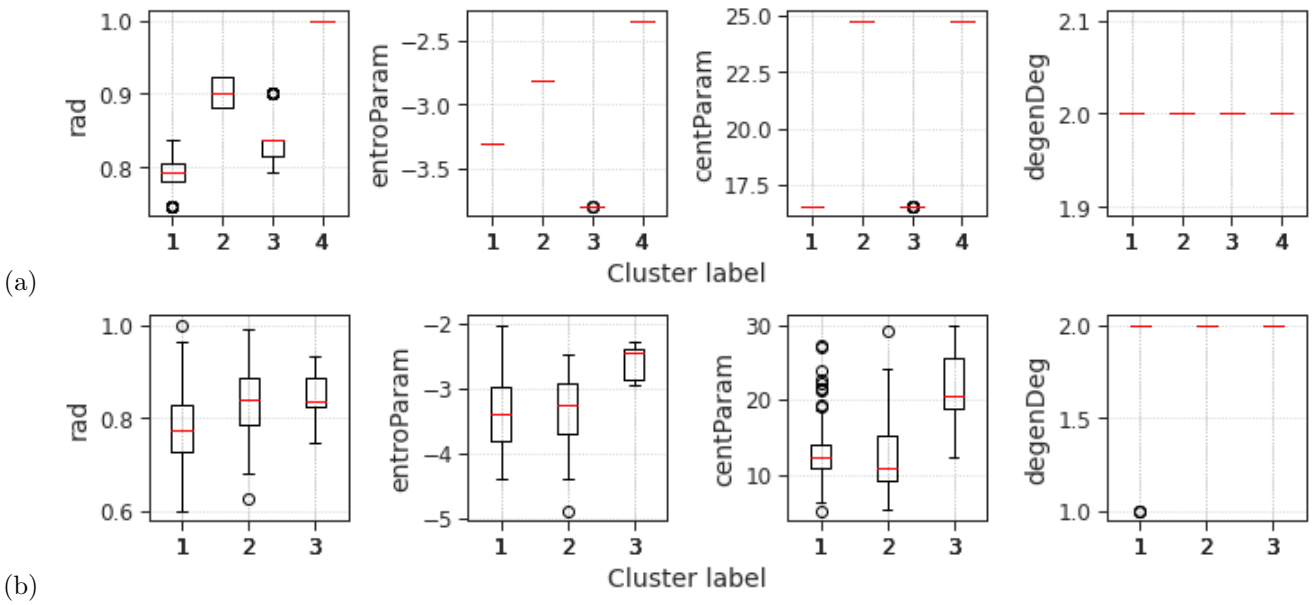


Figure S44: Box plots of radial distance from nanoparticle centre, entropy parameter, centrosymmetry parameter, and degeneracy degree for (a) the ordered cuboctahedron nanoparticle simulated at 0 K, and (b) the smallest disordered nanoparticle simulated at 723 K.

S15 Internal evaluation results

Table S3 tabulates the internal evaluation scores obtained from the clustering results for all nanoparticle data sets using the original feature space (all features included). Table S4 lists the scores for all disordered nanoparticle (not simulated at 0 K) data sets using the smaller set of features.

The results from Table S3 indicate that better partitions from the surface clustering are obtained for more ordered nanoparticles, as seen from the generally higher Silhouette scores and Calinski-Harabasz indices, and lower Davies-Bouldin indices. This is expected as the features used in this work are purely geometrical or structural. Structural disorder due to thermal energy in simulations inevitably introduces noises to the feature space, causing reduction in the differences between the surface atoms and therefore adding difficulties to the recognition of distinct surface patterns.

A comparison of the results from both tables also shows that the restriction of feature space to the smaller set of features chosen in Section S12 improves the clustering partitions.

Shape	T (K)	N_{total}	N_{surf}	Surface atoms			All atoms		
				S	CH	DB	S	CH	DB
CO	0	826	336	0.36	162.9	1.14	0.22	315.2	1.99
IC	0	923	362	0.55	485.5	0.72	0.27	394.0	1.38
TO	0	807	348	0.38	225.0	1.00	0.26	524.9	1.83
CU	0	1026	416	0.35	342.5	1.40	0.36	845.2	1.32
		1098	432	0.52	421.4	0.95	0.31	608.2	1.57
		1372	508	0.28	162.8	1.55	0.25	395.0	2.24
DH	0	686	297	0.36	105.4	1.13	0.20	207.4	1.73
		967	382	0.37	164.4	1.21	0.30	429.0	1.41
		1132	437	0.35	151.9	1.38	0.30	417.4	1.49
OT	0	670	326	0.41	156.9	1.28	0.48	815.2	1.41
		891	402	0.46	243.9	1.13	0.52	1265.6	1.30
		1156	486	0.26	213.3	1.39	0.48	1440.6	1.47
	323	670	326	0.25	72.9	1.93	0.47	677.6	0.90
		891	402	0.30	124.6	1.64	0.51	1084.7	0.84
		1156	486	0.28	106.2	1.77	0.51	1376.4	0.84
	523	670	326	-	-	-	0.27	287.0	1.58
		891	402	0.20	51.5	2.39	0.30	414.6	1.52
		1156	486	0.17	44.4	2.36	0.30	647.9	1.40
RD	0	1163	546	0.33	450.3	1.27	0.41	1285.2	1.27
		1524	662	0.29	511.4	1.46	0.37	1541.1	1.52
		1957	794	0.34	346.3	1.42	0.33	1359.9	1.73
	323	1163	524	0.21	89.5	1.68	0.45	437.9	1.18
		1524	654	0.29	155.9	1.38	0.41	571.8	1.55
		1957	775	0.28	183.4	1.34	0.41	642.5	1.54
	523	1163	489	0.27	99.1	1.61	0.23	351.8	3.33
		1524	604	0.12	72.9	2.05	0.25	395.5	1.77
		1957	726	0.00	24.2	4.84	0.37	559.0	1.79
TH	0	875	420	0.29	160.8	1.24	0.41	717.1	1.54
		969	514	0.22	134.3	1.80	0.34	471.6	1.91
		1330	650	0.23	159.0	1.84	0.15	438.3	2.51
	323	875	420	0.33	128.1	1.37	0.40	624.6	1.65
		969	512	0.22	41.8	2.04	0.35	284.4	1.82
		1330	648	0.33	82.5	1.45	0.46	1165.2	0.95
	523	875	421	0.25	65.0	1.73	0.42	673.6	1.04
		969	475	0.12	20.4	2.11	0.36	324.1	1.30
		1330	622	-0.08	20.5	2.17	0.37	527.9	1.67
DIS	723	641	290	0.03	7.6	4.57	0.15	195.4	2.29
		864	366	-0.07	7.9	3.46	-0.02	131.3	3.35
		1098	436	0.17	40.1	1.88	0.16	378.9	2.95

Table S3: Internal evaluation scores of the clustering of the surface and all atoms using feature space consisting of all features for all nanoparticles, indicating the temperature (T), the total number of atoms (N_{total}), and number of surface atoms (N_{surf}). Good clustering partitions are indicated by higher Silhouette (S) scores (in the range of $[-1, 1]$), higher Calinski-Harabasz (CH) indices (in the range of $[0, \infty)$), and lower Davies-Bouldin (DB) indices (in the range of $[0, \infty)$). No score can be computed for nanoparticles with only one surface atom cluster.

Shape	T (K)	N_{total}	N_{surf}	Surface atoms			All atoms		
				S	CH	DB	S	CH	DB
OT	323	670	326	0.37	160.1	1.12	0.69	2025.2	0.48
		891	402	0.43	234.1	1.05	0.52	1329.6	1.19
		1156	486	0.49	338.2	1.01	0.32	1988.5	1.42
	523	670	326	0.32	114.7	1.48	0.65	1794.7	0.54
		891	402	0.39	181.5	1.28	0.65	2381.1	0.54
		1156	486	0.38	179.8	1.46	0.54	2182.8	1.30
RD	323	1163	524	0.26	161.3	1.63	0.54	1649.4	1.07
		1524	654	0.31	197.4	1.31	0.43	1686.1	1.03
		1957	775	0.21	256.0	1.68	0.51	1892.8	1.68
	523	1163	489	0.17	77.5	1.96	0.48	1482.4	1.28
		1524	604	0.13	90.6	2.19	0.45	865.4	1.55
		1957	726	0.08	77.0	2.09	0.46	1204.2	1.87
TH	323	875	420	0.59	395.0	0.74	0.39	1536.0	1.08
		969	512	0.36	255.8	1.22	0.66	1758.7	0.95
		1330	648	0.42	329.2	1.23	0.67	3184.0	0.69
	523	875	421	0.44	283.7	1.09	0.46	1163.7	1.03
		969	475	0.18	58.8	1.66	0.50	1094.9	0.71
		1330	622	0.17	60.6	1.89	0.53	1464.8	0.91
DIS	723	641	290	0.10	26.1	2.11	0.49	832.9	0.84
		864	366	0.06	33.4	2.45	0.47	973.4	0.89
		1098	436	0.02	21.5	2.82	0.30	917.7	1.41

Table S4: Internal evaluation scores of the clustering of the surface and all atoms using feature space consisting of the Steinhardt, neighbour, and order descriptors for all disordered nanoparticles, indicating the temperature (T), the total number of atoms (N_{total}), and number of surface atoms (N_{surf}). Good clustering partitions are indicated by higher Silhouette (S) scores (in the range of $[-1, 1]$), higher Calinski-Harabasz (CH) indices (in the range of $[0, \infty)$), and lower Davies-Bouldin (DB) indices (in the range of $[0, \infty)$).

S16 Visualisation of domain relevant cluster evaluation

Figures S45 and S46 show the comparisons between the catalytic weighting profiles obtained from the reference activity maps for the chemical reactions and the surface cluster GCN distributions for both nanoparticles, from which the selectivity, specificity, and sensitivity scores were computed.

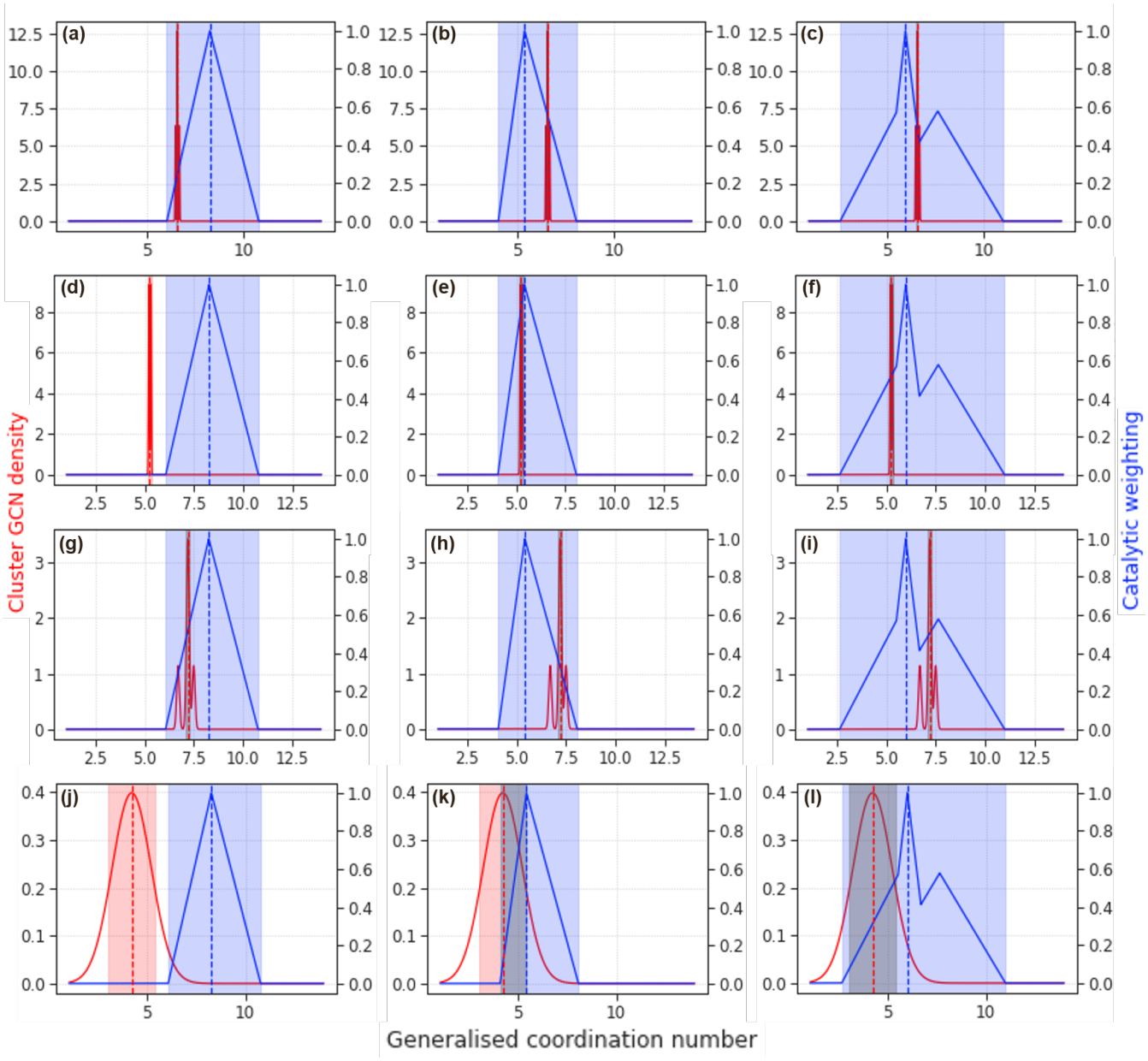


Figure S45: Evaluation of the generalised coordination number distributions of clusters 1 (a-c), 2 (d-f), 3 (g-i), and 4 (j-l) of an ordered cuboctahedron nanoparticle simulated at 0 K by the catalytic weighting towards oxygen reduction reaction (a, d, g, j), carbon monoxide oxidation reaction (b, e, h, k), aliphatic ketone reduction reaction (c, f, i, l).

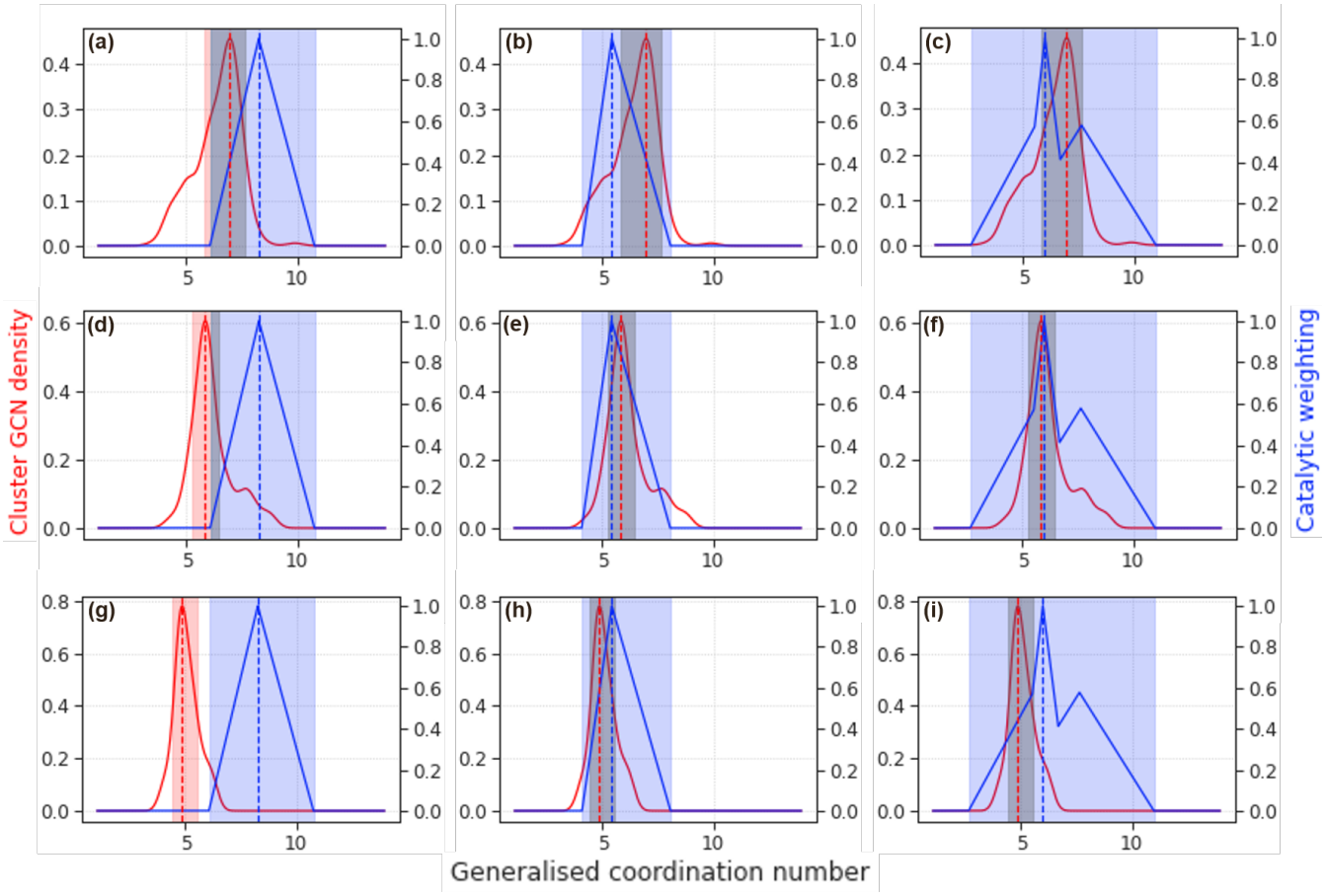


Figure S46: Evaluation of the generalised coordination number distributions of clusters 1 (a-c), 2 (d-f), and 3 (g-i) of the smallest disordered nanoparticle simulated at 723 K by the catalytic weighting towards oxygen reduction reaction (a, d, g) , carbon monoxide oxidation reaction (b, e, h), aliphatic ketone reduction reaction (c, f, i).

References

[1] G. J. Ackland and A. P. Jones, *Physical Review B*, 2006, **73**, 054104.

[2] P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Physical Review B*, 1983, **28**, 784–805.

[3] C. L. Kelchner, S. J. Plimpton and J. C. Hamilton, *Physical Review B*, 1998, **58**, 11085–11088.

[4] P. M. Piaggi, O. Valsson and M. Parrinello, *Physical Review Letters*, 2017, **119**, 15701.

[5] G. Opletal, G. Grochola, Y. H. Chui, I. K. Snook and S. P. Russo, *Journal of Physical Chemistry C*, 2011, **115**, 4375–4380.

[6] A. J. Parker, G. Opletal and A. S. Barnard, *Journal of Applied Physics*, 2020, **128**, 1–11.

[7] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.

[8] S. Menon, G. D. Leines and J. Rogal, *Journal of Open Source Software*, 2019, **4**, 1824.

[9] T. Pilati and A. Forni, *Journal of Applied Crystallography*, 1998, **31**, 503–504.

[10] W. Mickel, S. C. Kapfer, G. E. Schröder-Turk and K. Mecke, *Journal of Chemical Physics*, 2013, **138**, 44501.

[11] W. Lechner and C. Dellago, *The Journal of chemical physics*, 2008, **129**, 114707.

[12] J. A. van Meel, L. Filion, C. Valeriani and D. Frenkel, *Journal of Chemical Physics*, 2012, **136**, 234107.

[13] O. J. Murphy and S. M. Selkow, *Information Processing Letters*, 1990, **34**, 37–41.

[14] A. Stukowski, V. V. Bulatov and A. Arsenlis, *Modelling and Simulation in Materials Science and Engineering*, 2012, **20**, 85007.

[15] T. Schilling, H. J. Schöpe, M. Oettel, G. Opletal and I. Snook, *Physical Review Letters*, 2010, **105**, 025701.

[16] T. Schilling, S. Dorosz, H. J. Schöpe and G. Opletal, *Journal of Physics: Condensed Matter*, 2011, **23**, 194120.

[17] A. Stukowski, *Modelling and Simulation in Materials Science and Engineering*, 2012, **20**, 045021.

[18] P. M. Larsen, *arXiv*, 2020, **arXiv:2003.08879**, 1–11.

[19] G. Opletal, J. Y. C. Ting and A. S. Barnard, *NCPac*, 2024, `https://doi.org/10.25919/tfv3-he58`.

[20] E. A. Rakhmanov, E. B. Saff and Y. M. Zhou, *Mathematical Research Letters*, 1994, **1**, 647–662.

[21] L. V. D. Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

[22] D. Kobak and G. C. Linderman, *Nature Biotechnology 2021 39:2*, 2021, **39**, 156–157.

[23] R. Gove, L. Cadalzo, N. Leiby, J. M. Singer and A. Zaitzeff, *Visual Informatics*, 2022, **6**, 87–97.

[24] A. Rodriguez and A. Laio, *Science*, 2014, **344**, 1492–1496.

[25] G. Palshikar, Proceeding of the 1st International Conference on Advanced Data Analysis, Business Analytics and Intelligence, 2009, pp. 1–13.

[26] P. J. Rousseeuw, *Journal of Computational and Applied Mathematics*, 1987, **20**, 53–65.

[27] T. Caliński and J. Harabasz, *Communications in Statistics*, 1974, **3**, 1–27.

[28] D. L. Davies and D. W. Bouldin, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, **PAMI-1**, 224–227.

[29] F. Calle-Vallejo, J. Tymoczko, V. Colic, Q. H. Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann and A. S. Bandarenka, *Science*, 2015, **350**, 185–189.

[30] F. Calle-Vallejo, M. D. Pohl and A. S. Bandarenka, *ACS Catalysis*, 2017, **7**, 4355–4359.

[31] M. Hu, J. He, R. Guo, W. Yuan, W. Xi, J. Luo and Y. Ding, *Catalysis Communications*, 2020, **146**, 106129.

[32] Z. Zhao, Z. Chen, X. Zhang and G. Lu, *Journal of Physical Chemistry C*, 2016, **120**, 28125–28130.

[33] C. J. Bondue, F. Calle-Vallejo, M. C. Figueiredo and M. T. M. Koper, *Nature Catalysis*, 2019, **2**, 243–250.

[34] Y. Kalambet, Y. Kozmin and A. Samokhin, *Chemometrics and Intelligent Laboratory Systems*, 2018, **179**, 22–30.