SUPPORTING INFORMATION

Jonathan Y. C. Ting [*1] and Amanda S. Barnard [†1]

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

S1 Archetypal Analysis Acceleration

Archetypal analysis is a powerful method that has yet to be fully leveraged in science, due in part to computational (rather than statistical) challenges. The inherent complexity of AA has been reduced *via* sparse representations[1,2] and usage of coresets[3]. Sparse representations of the original data downsizes the AA problem prior to the minimisation of Eqn. 1 in main text by reducing the number of samples. The representations allow the convex hull of $\mathbf{X}$ to be approximated by with a polytope with fewer extreme points, enabling faster AA. They can be obtained *via* either random projections[1] or non-negative least squares[2]. Han *et al.* combined random projection with another preprocessing technique known as randomised low-rank approximation to reduce the number of dimensions by replacing $\mathbf{X}$ with a low-dimensional representation $\widetilde{\mathbf{X}}$[4]. The combined method effectively reduces the scaling of AA, provided the data $\mathbf{X}$ is approximated well by the embedding in low-dimensional linear subspace, and the convex hull of $\mathbf{X}$ is well approximated by a polytope with fewer vertices. While these techniques can be combined with other AA acceleration techniques, they are not employed in this work because it is unknown whether our testbed datasets satisfy these conditions.

Careful initialisation of the coefficients is known to improve the speed of convergence and lowers the risk of finding insignificant archetypes, in particular to select mixtures that maintain sufficient separation from each other. An approach that addresses the steps the prior to the alternating minimisation step are improved initialisation procedures and approximation of the original data. Initialisation procedures such as `FurthestSum`[5] and `AA++`[6] have been proposed. The `FurthestSum` approach is inspired by the `FurthestFirst` approach widely used for k-means[7], while `AA++` is a probabilistic initialisation strategy that sequentially samples samples based on their influence on the objective, similar to k-`means++`[6].

Coresets are compact, weighted subsets of data that approximates the original dataset, such that the performance of models applied to the subset is provably competitive compared

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

to operations on the whole dataset[8]. Provided that the coresets approximate the full dataset sufficiently well, AA can be conducted using the coresets with similar performance as the full dataset within shorter execution time. Mair *et al.* introduced an efficient coreset construction algorithm that requires only two passes over the data[3]. The authors managed to conduct theoretical quantification of the approximation error, ensuring that the performance on the coreset is competitive with the performance on the full dataset. Coreset is deemed to be a theoretically sound alternative to the other approaches[3].

IAA resembles the coreset method[3] in that a subset of samples is extracted from the original dataset, and hence the overall AA is approximate in nature. It differs in that the samples extracted from AA on subsets of the original data might or might not exist in the original dataset (hypothetical in nature), and do not (yet) qualify as coresets in terms of mathematical provability. With only two passes over the whole dataset required to identify the coresets, the coreset method is arguably more computationally efficient than IAA. However, the highly parallelisable nature of IAA can be faster in execution for the subset identification step, especially when the dataset is enormous. IAA is not related to the geometric approach proposed by Abrol and Sharma[9] despite the similar claim of both approaches in leveraging the underlying geometry of AA. Additionally, the common usage of the term "iterative" in the namings (the geometric AA employs an iterative approach to subset selection) also corresponds to different steps of the AA problem.

S2 Conventional Domain-Driven Sample Reduction

As mentioned in the main text, the domain-driven approach to identify and remove redundant conformations is applied during the sampling from MD simulations. The reduction workflow is illustrated in Fig. S1.

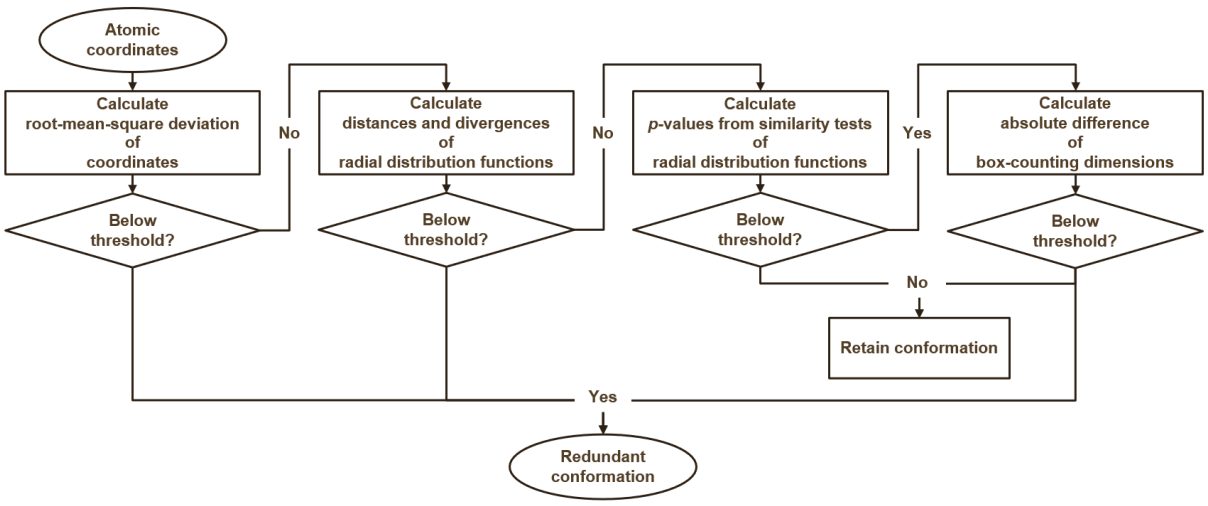


Figure S1: Overview of the workflow of the identification of redundant nanoparticle conformations.

The threshold of each similarity metric is decided by observing the changes of the metric during the melting simulation trajectories for a few chosen nanoparticles. For each given pair of conformations, the first conformation is considered redundant if and only if:

1. the coordinates of the pair has an root mean squared error smaller than a threshold (set to 0.05 Å),

2. the average bond length distributions of the pair has mutual information[10,11] greater than a threshold (set to 1.0),

3. the p values from 2-sample Anderson-Darling test[12], 2-sample Kolmogorov-Smirnov test[13], and 2-sample Cramér-von Mises test[14] on the pair of radial distribution functions are greater than a threshold (set to the standard level of significance at 0.05),

4. the Hellinger distance, Jensen-Shannon divergence and Kantorovich-Rubinstein metric between the radial distribution functions of the pair are less than their thresholds (set to 0.08, 0.005, and 0.0001, respectively), and

5. the Kullback-Leibler divergence from the radial distribution functions of the first conformation to second conformation is less than a threshold (set to 0.25) (note that the metric is not symmetric, *i.e.* $KL(P,Q)! = KL(Q,P)$)

The samples in the bimetallic datasets are reduced using a smaller selection of similarity metrics, based on the consideration of computational resources and amount of information captured. Two metrics are excluded from the sample reduction pipeline, namely the mutual information of the distribution of average bond length of each atom, and Kullback-Leibler divergence of the radial distribution function (RDF) of current conformation from previous conformation. The threshold for each similarity metric is listed below:

1. root-mean-square deviation (calculated in Euclidean distance) of the current conformation from the previous conformation (with a threshold of 1.0),

2. Wasserstein distance[15] (with a threshold of 0.0002), Hellinger distance[16] (with a threshold of 0.15), and Jensen-Shannon divergence[17] of the RDF of the current conformation from the RDF of the previous conformation (with a threshold of 0.02), and

3. p-values (with thresholds of 0.05) from 2-sample Anderson-Darling test, 2-sample Kolmogorov-Smirnov test, and 2-sample Cramer-von Mises test, which test the hypothesis that there is no significant difference between the radial distribution functions of the current and previous conformations.

Data set	Number of features	Original Number of instances	Minimum and maximum atom number	Minimum and maximum diameter (nm)	Reference to public repository	Final Number of instances[a]	Percentage of reduction (%)
Au	182	4000	236-14277	1.7-7.8	[18]	-	-
Pd	182	4000	137-16262	1.4-7.5	[19]	-	-
Pt	182	1300	54-15837	1.5-7.6	[20]	-	-
AuPd	922	145103	93-4631	1.1-5.1	[21]	47623	67.2
AuPt	922	162439	93-4631	1.1-5.1	[22]	53445	67.1
PdAu	922	145064	105-4631	1.2-5.0	[23]	48087	66.9
PdPt	922	151216	105-4631	1.2-5.0	[24]	48785	67.7
PtAu	922	162770	105-4631	1.2-5.1	[25]	54006	66.8
PtPd	922	150781	105-4631	1.2-5.0	[26]	48943	67.5
AuPdPt	1958	48136	603-959	2.3-2.9	[27]	-	-

[a] Only instances from bimetallic nanoparticles are reduced, with the domain-driven methods described in Section S2. The range of the number of atoms and diameters remained the same after the reduction.

Table S1: Descriptions of the metal nanoparticle data sets used in the main text.

S3 Metal Nanoparticle Datasets

S4 IAA Parameters for Metal Nanoparticle Datasets

Table S2 lists all IAA parameters used for sample redundancy reduction of metal nanoparticle datasets in the main text. The numbers of subsets are chosen to ensure that each subset is sufficiently large to return the specified number of archetypes during the elbow plot analysis (both number of intermediate and final archetypes, p_1 and p_2, were kept constant at this stage). Taking monometallic as example, each subset has to be larger than 300 archetypes (which is the maximum number explored in the elbow plot analysis). Given that it has a total of 9300 samples, splitting the monometallic dataset into 100 subsets with 93 samples each is inappropriate. For the IAA runs with the maximum number (p_2=5000) of archetypes and with no redundant archetypes, the largest number of intermediate archetypes (p_1) are used, as informed by the results from Section S5, which indicate that larger p_1 helps in preserving the explained variance from the AA on the subsets.

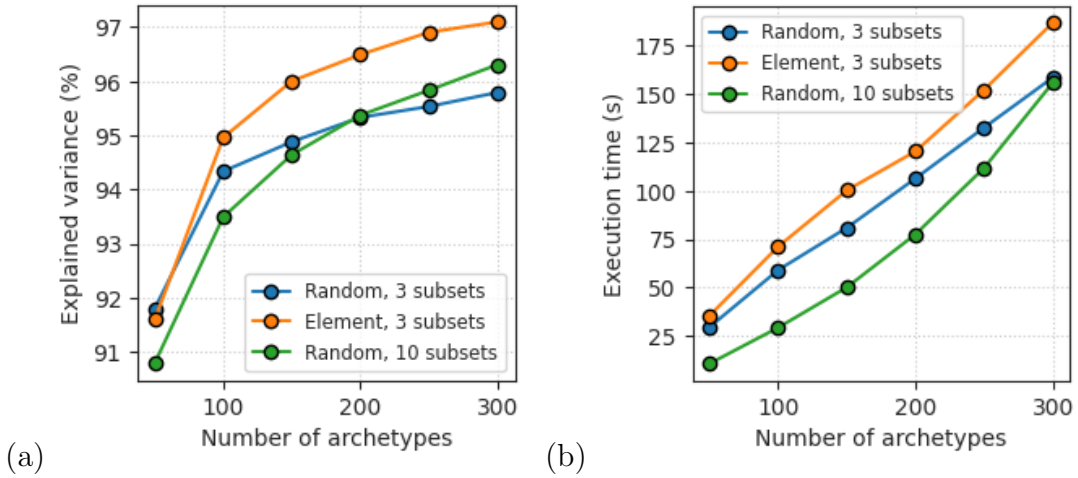
S5 Comparisons between Random and Element-Based Splitting

There are multiple ways to split the original data into subsets. Domain knowledge-informed splitting based on the elemental composition of the nanoparticles is compared to random

IAA runs with	Parameters	Monometallic	Bimetallic	Trimetallic
Elbow plot-informed archetypes	k	10	100	10
	p_1	100	200	200
	p_2	100	200	200
Maximum number of archetypes	k	10	100	10
	p_1	300	500	1000
	p_2	5000	5000	5000
No redundant archetype	k	10	100	10
	p_1	300	500	1000
	p_2	250	1800	3000

Table S2: Parameters of iterative archetypal analysis for the monometallic, bimetallic, and trimetallic nanoparticle datasets, including the number of subsets (k), number of intermediate archetypes (p_1), and number of final archetypes (p_2).

splitting, while fixing the other parameters. The numbers of archetypes for the first and second rounds of AA in the IAA procedure (p_1 and p_2) are set to be the same for this example. Fig. S2 shows that the domain knowledge-based splitting results in higher explained variance than random splitting. For random splitting, using more subsets also improved the amount of explained variances when more than 200 archetypes are returned. When the original data is split into more subsets, there is higher chance for the subsets to be distributed in ways that require larger number of archetypes to capture the variance of the subsets.



(a) (b)

Figure S2: (a) Explained variance from iterative archetypal analysis with different number of archetypes, with subset selection *via* either random or element-based splitting, and (b) the execution time.

To visualise the high-dimensional data in two dimensional figures, the t-SNE method implemented in the `scikit-learn` package is used, with `n_components`=2, `perplexity`=30.0,

`early_exaggeration`=12.0, `learning_rate`='auto', `n_iter`=1000, `n_iter_without_progress`=300, `min_grad_norm`=1×10^{-7}, `metric`=metric, `random_state`=42, `init`='pca', `method`='barnes_hut', and `angle`=0.5. A comparison between the distributions of the archetypes obtained from random and element-based splitting in Fig. S3 shows that the element-based splitting approach tends to return less archetypes in the lower left region of the 2D embedded space from *t*-SNE. Judging from the higher explained variance from the element-based splitting approach, a large proportion of the samples in the region are likely similar to each other and hence can be explained sufficiently well by a small number of archetypes. This allows the remaining archetypes extracted from the element-based splitting approach to focus on explaining the variance of the samples in the other regions.

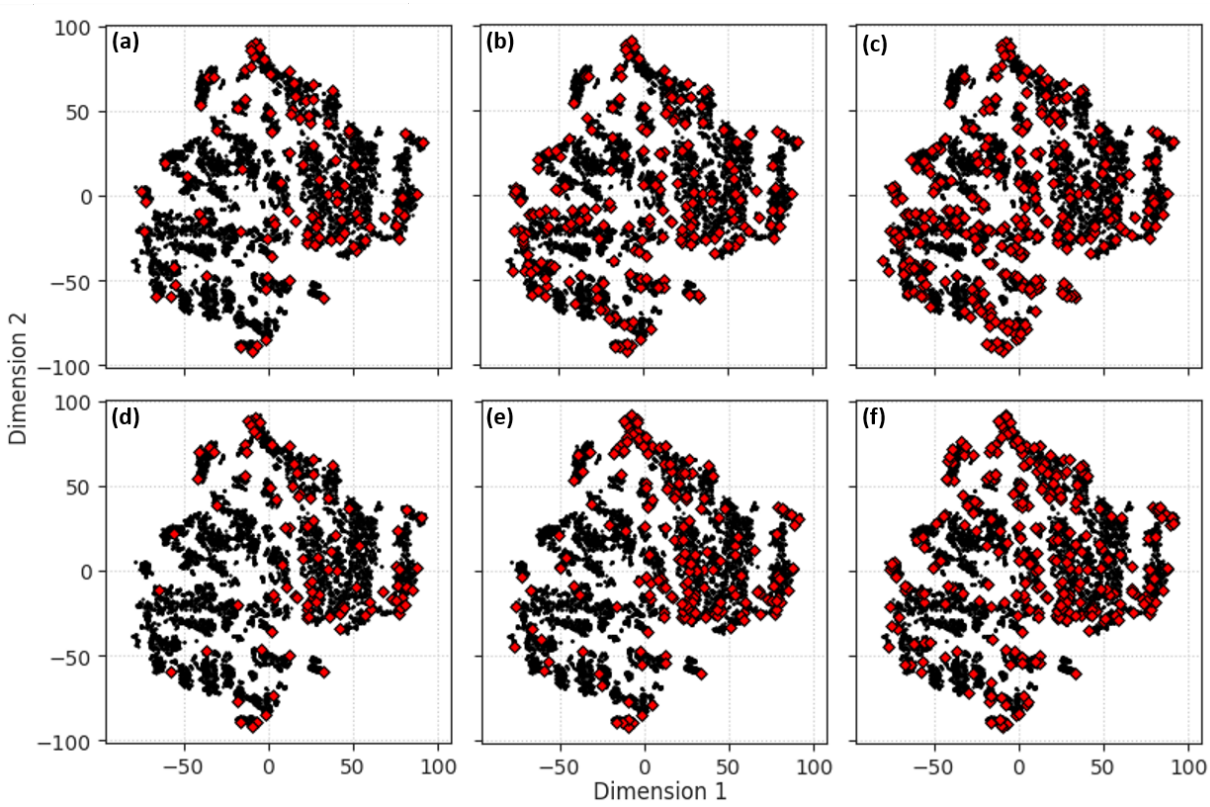


Figure S3: Mapping of the samples (black dots) and archetypes (red stars) of the monometallic nanoparticle dataset onto 2D embedded space learnt *via* *t*-distributed stochastic neighbour embedding, with (a, d) 100, (b, e) 200, and (c, f) 300 archetypes obtained from (a-c) random and (d-f) element-based splitting with 3 subsets. The reduced dimensions are arbitrary.

While the element-based splitting is found to be superior over random splitting, it does not necessarily mean that domain knowledge-informed splitting is always better. Element-based splitting cannot be applied to datasets that are comprised of single elemental combination such as the trimetallic dataset, and splittings based on continuous variables such

as sizes would require a potentially arbitrary threshold to define the classes. Therefore, random splitting is still employed for all metal nanoparticle datasets.

Fig. S4, S5, and S6 show the distributions of the samples and archetypes obtained from different subset splitting methods and different number of subsets on the monometallic dataset.

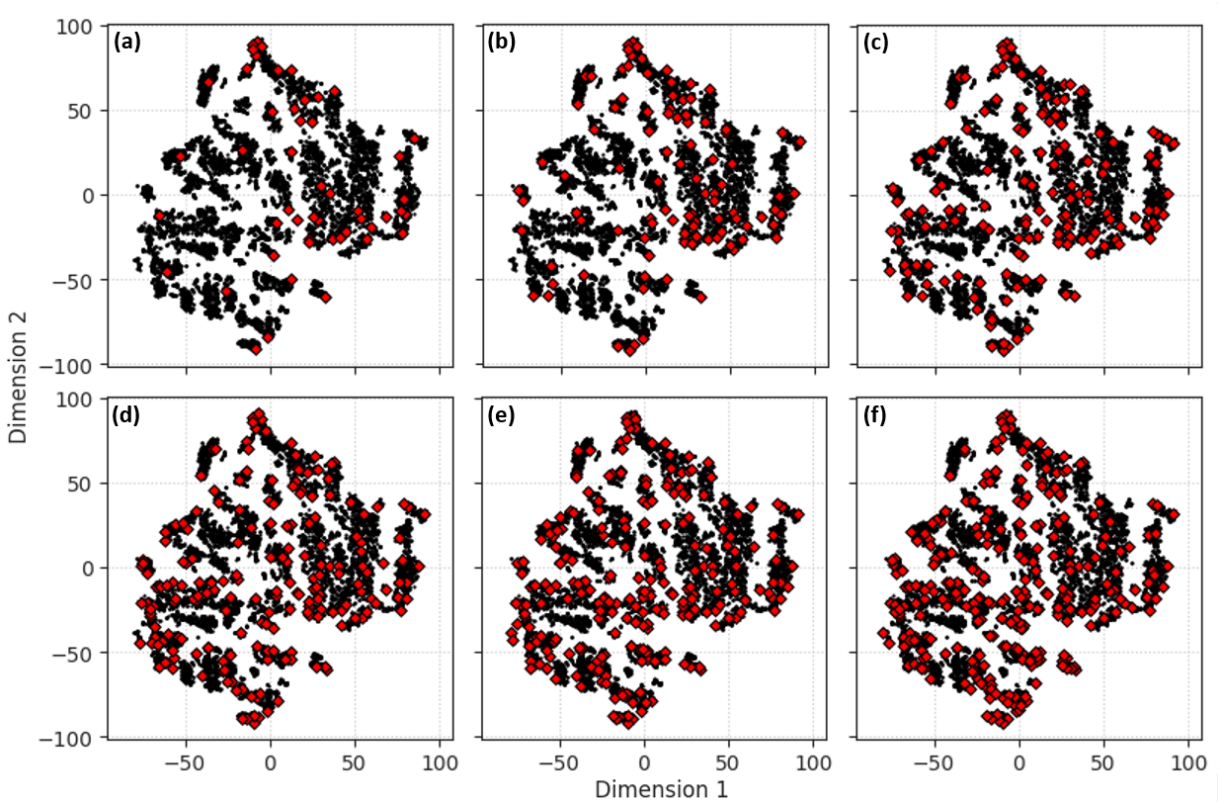


Figure S4: Mapping of the samples (black dots) and archetypes (red stars) obtained from random splitting of the monometallic nanoparticle dataset into 3 subsets onto 2D embedded space learnt *via* t-distributed stochastic neighbour embedding. The reduced dimensions are arbitrary.

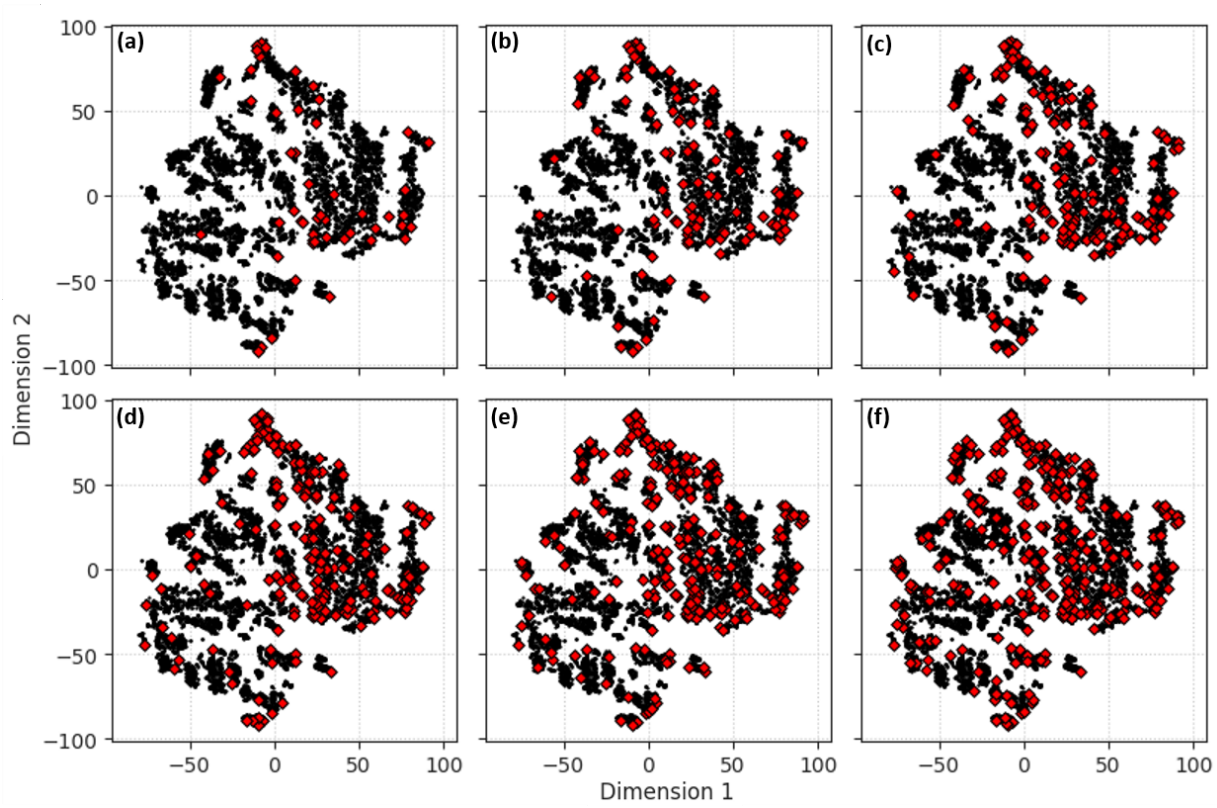


Figure S5: Mapping of the samples (black dots) and archetypes (red stars) obtained from element-based splitting of the monometallic nanoparticle dataset into 3 subsets onto 2D embedded space learnt *via* t-distributed stochastic neighbour embedding. The reduced dimensions are arbitrary.

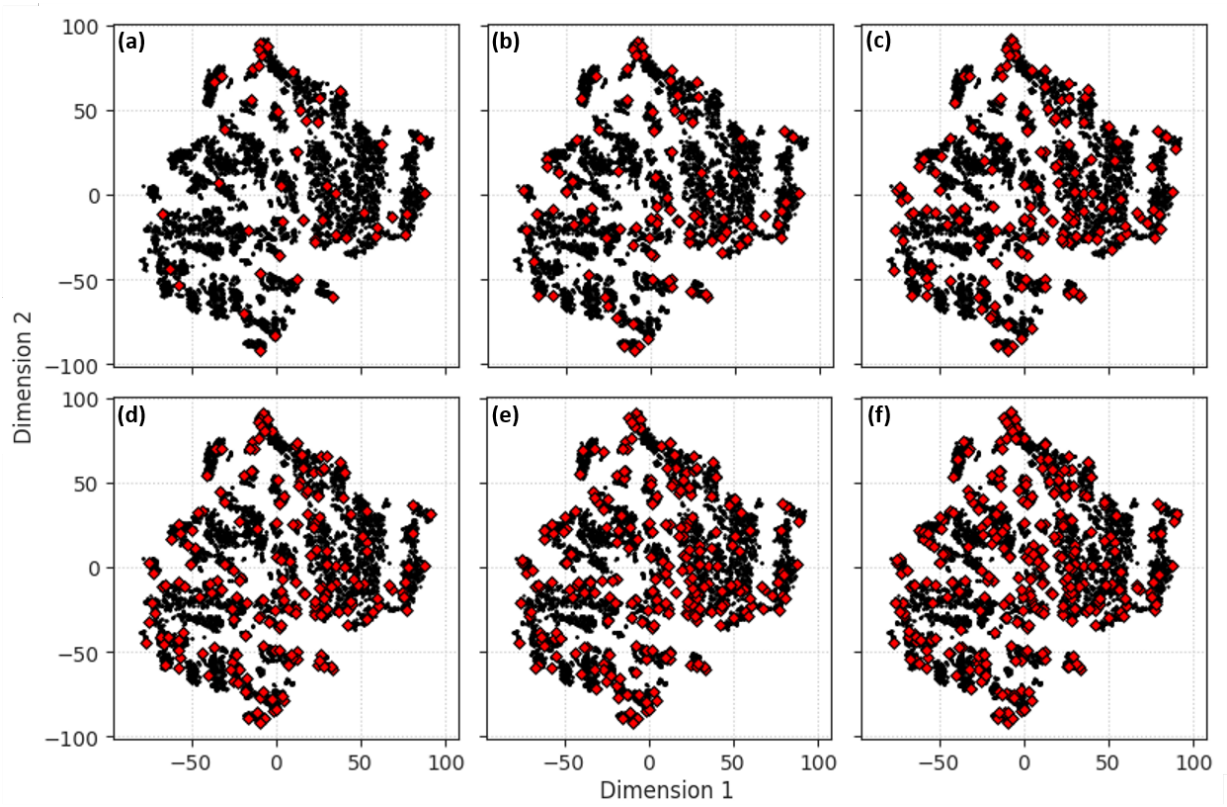


Figure S6: Mapping of the samples (black dots) and archetypes (red stars) obtained from random splitting of the monometallic nanoparticle dataset into 10 subsets onto 2D embedded space learnt *via* t-distributed stochastic neighbour embedding. The reduced dimensions are arbitrary.

S6 Optimising the Numbers of Archetypes

As mentioned in the main text, in this study the optimal number of archetypes is determined by gradually reducing the number of archetypes from the number known to result in redundant sampling (5000 archetypes), until the vertices of the simplex plots are completely occupied (or very closely populated) by samples that actually exist in the original dataset. Examples are shown in Fig. S7. It is noted that Fig. S7(d) has arguably more central prototype than Fig. S7(e) and (f), hence the centrality of the prototype's position alone does not guarantee the absence of redundant archetypes.

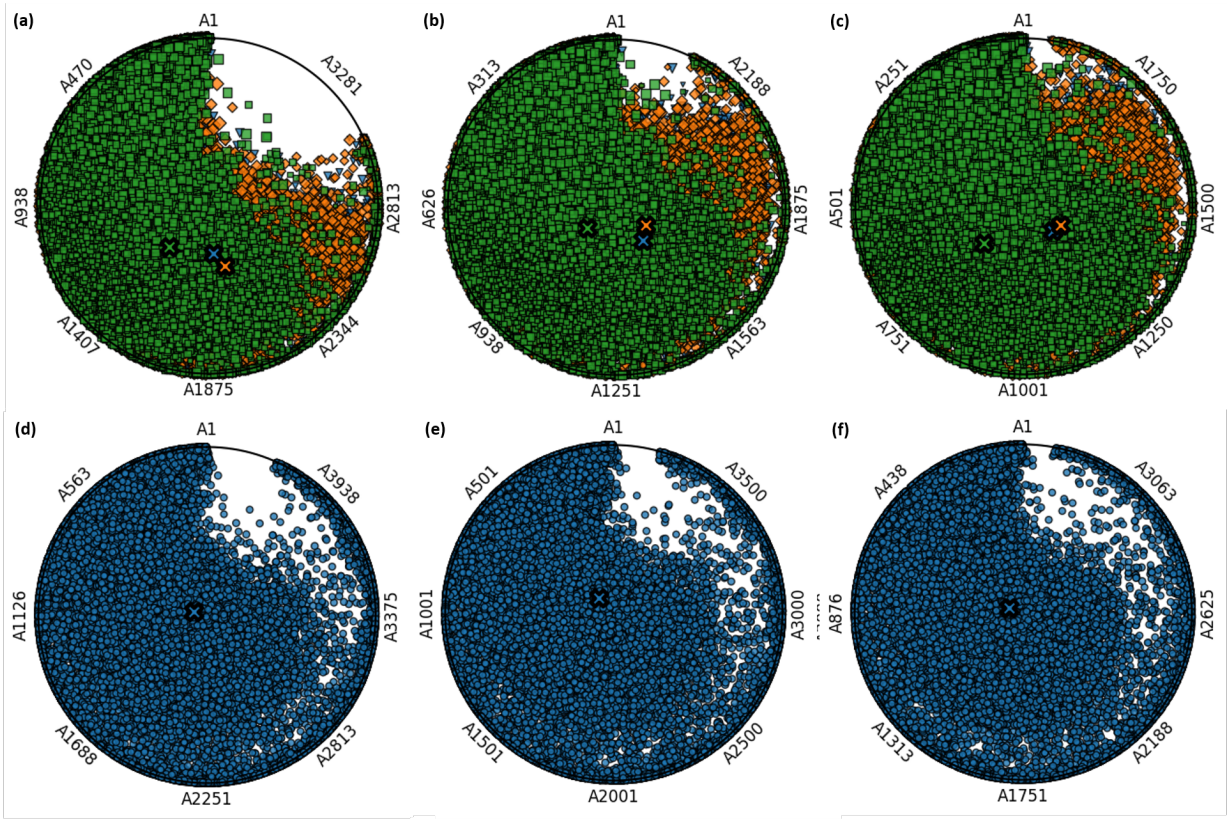


Figure S7: Simplex plots of the samples for bimetallic AuPd (blue), AuPt (orange), and PdPt (green) nanoparticle datasets with (a) 5000, (b) 2500, and (c) 2000 archetypes, and trimetallic AuPdPt (blue) nanoparticle dataset with (d) 5000, (e) 3750, and (f) 3500 archetypes. The crosses indicate the hypothetical prototypes with averaged feature values and the same elemental combinations. Points are sized relative to the average diameters of the nanoparticles.

S7 Contribution from Archetypes to Data Samples

The samples that are closest to the archetypes that lie within and outside the empty sectors are compared in Fig. S8. It is clearly seen that the archetypes that lie outside the mainly empty sectors contribute much more significantly to the closest samples, while the archetypes within the empty sectors contribute very little even to the samples closest to them. These observations indicate that choosing 5000 archetypes results in a poor sampling of the original datasets. The radar plots displaying the contribution of archetypes toward 100 randomly chosen samples shown in Fig. S9 also provided strong evidence that the later archetypes do not contribute significantly to any sample.

A possible explanation for the insignificant contributions of the later archetypes toward the reconstruction of the samples is that they are very similar to the archetypes that have already been returned. It is known that the commonly used `FurthestSum` initialisation procedure[28] is prone to selecting redundant archetypes (which lie on the convex hull of the already selected archetypes) when many archetypes are requested[6]. This was attributed to its early focus on suboptimal boundary points, which traps the results in poor local minima. Even though the 3D spatial structures of the hypothetical archetypes cannot be visualised, the feature profiles of the archetypes can be studied to verify that this is the case. As shown in Fig. S10, the profile of all features for the later archetypes that reside within the empty sectors are indeed visually indistinguishable feature profiles with each other, confirming their redundancy.

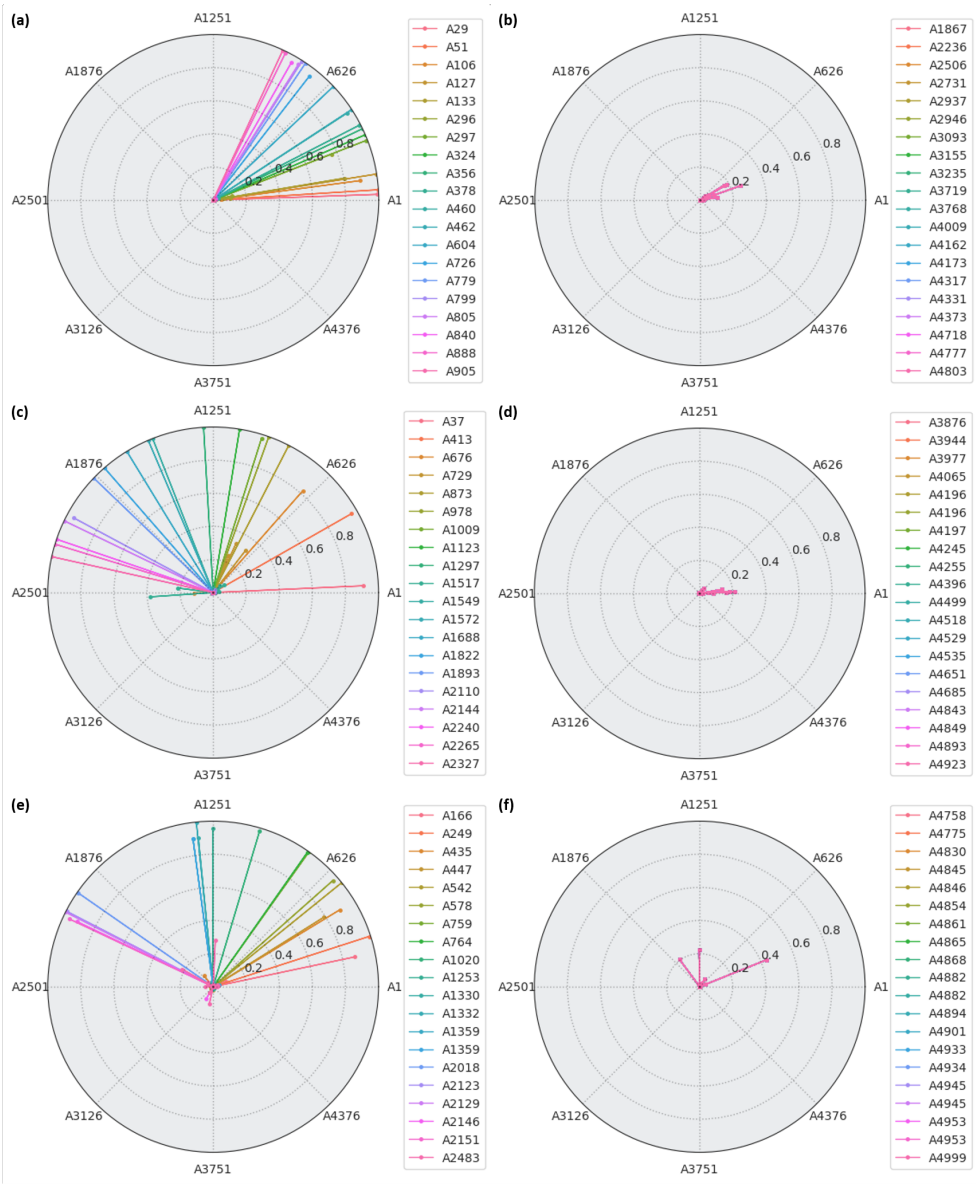


Figure S8: Radar plots indicating the contributions of all archetypes toward the samples that closely match the archetypes sampled from (a,c,e) those that lie outside the mainly empty sectors, and (b,d,f) those that lie within, for the (a-b) monometallic, (c-d) bimetallic, and (e-f) trimetallic nanoparticle datasets. "A" represents archetype.

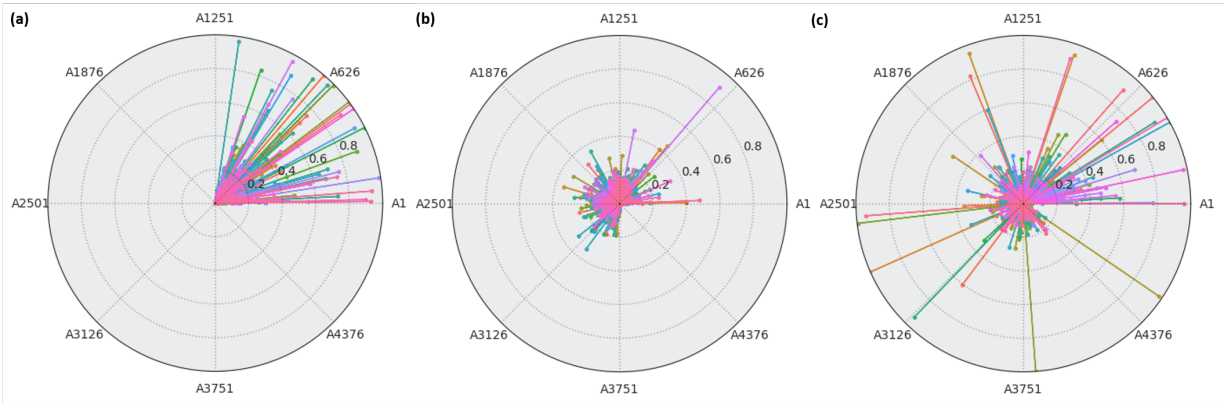


Figure S9: Radar plots indicating the contributions of all archetypes toward 100 randomly sampled samples for the (a) monometallic, (b) bimetallic, and (c) trimetallic nanoparticle datasets. "A" represents archetype. The colours of lines indicate the order of the archetypes (the HSL colour palette of `seaborn` Python package is used, refer to Fig. S8 for colour sequence).

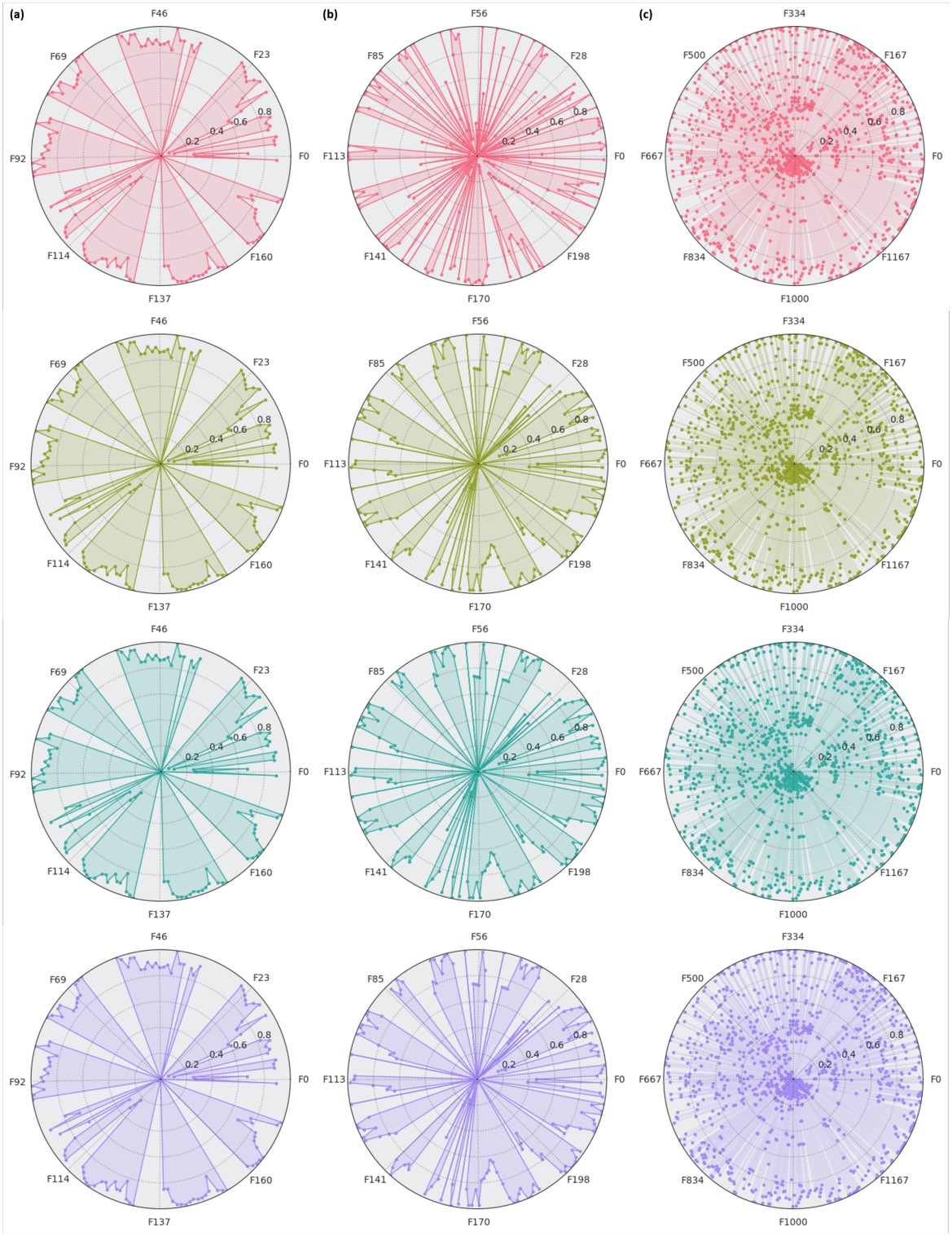


Figure S10: Radar plots displaying the feature profiles of archetypes from the mainly empty sectors of the simplex plots of the (a) monometallic, (b) bimetallic, and (c) trimetallic nanoparticle datasets.

S8 Archetype Characterisation

Visualisation of the feature profiles of the metal nanoparticle archetype can be displayed as radar and bar plots, as shown in Fig. S11 and S12. The radar plots are particularly useful for a comprehensive comparison of all features between the chosen archetypes, while a smaller set of features can then be compared using the bar plots.

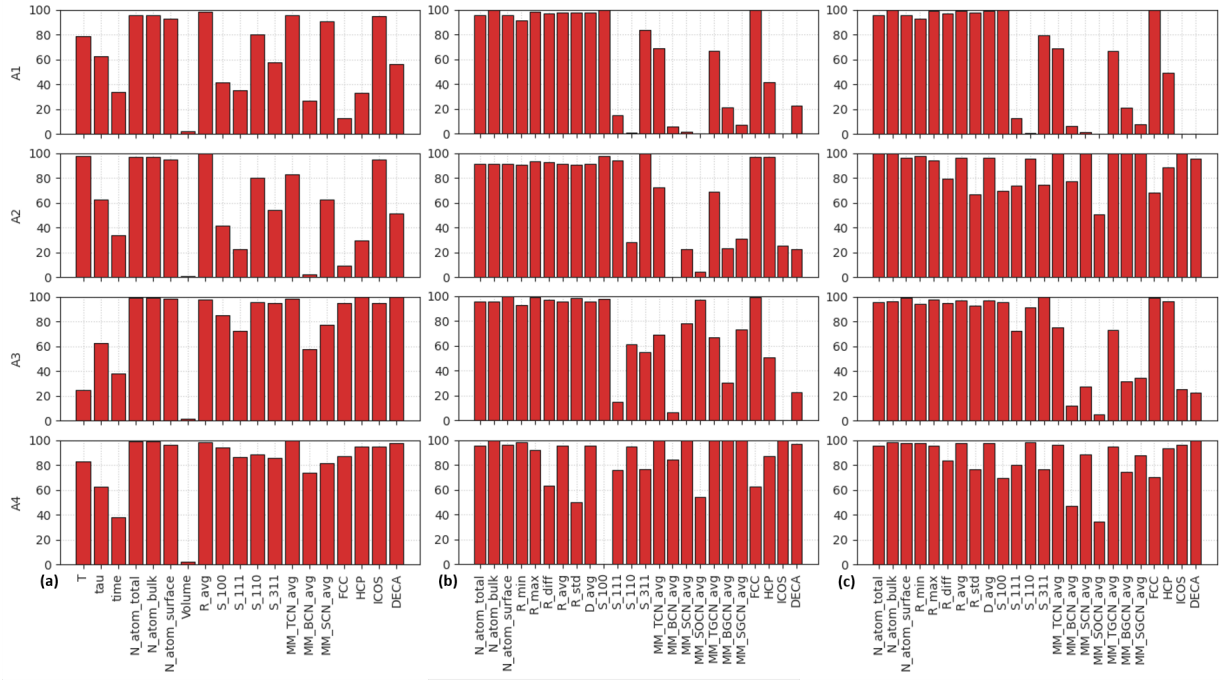


Figure S11: Bar plots of feature profiles for archetypes 1-4 from archetypal analysis of (a) monometallic, (b) bimetallic, and (c) trimetallic nanoparticle datasets. "A" represents archetype.

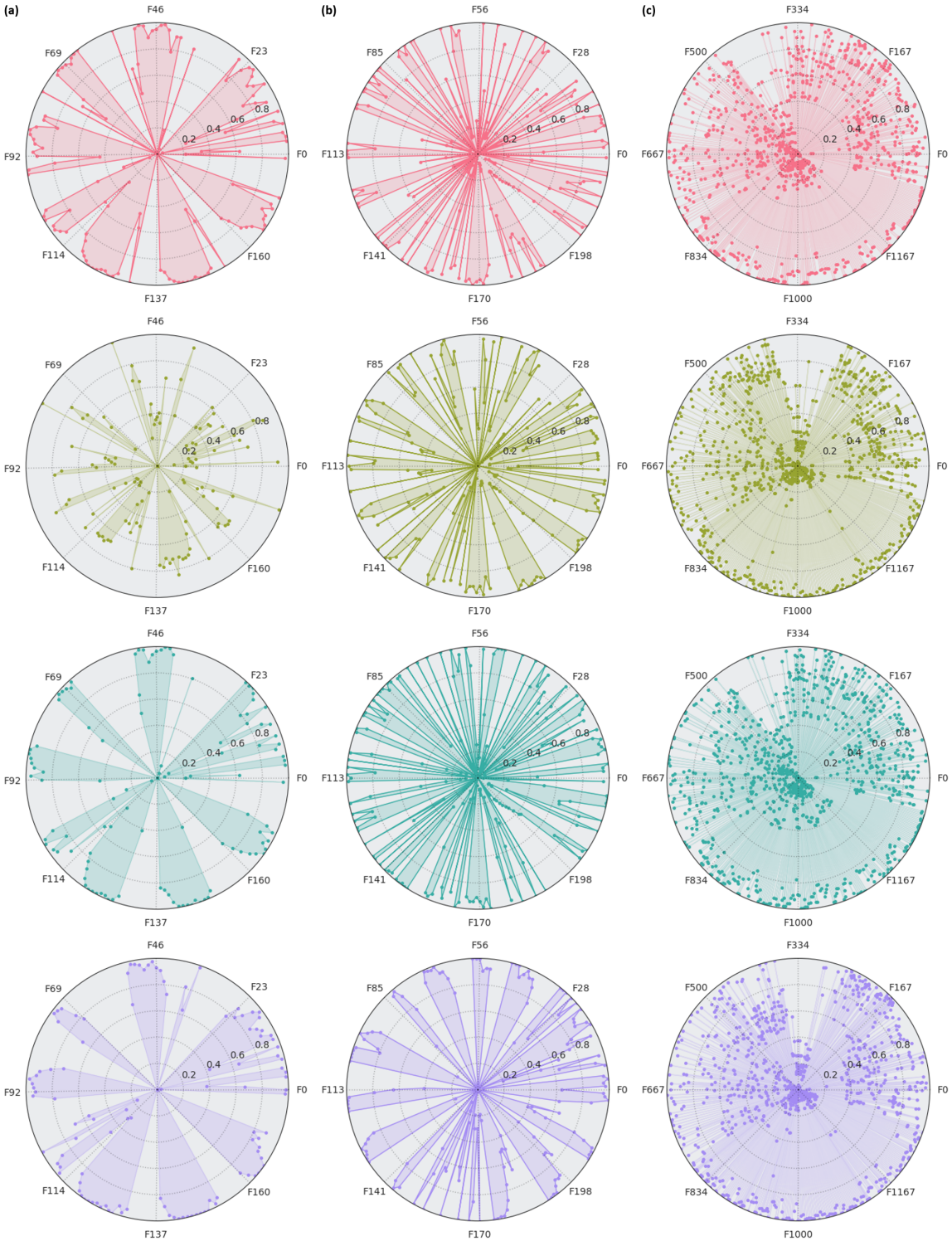


Figure S12: Radar plots of feature profiles for archetypes 1 (red), 2 (yellow), 3 (turquoise), and 4 (purple) from archetypal analysis of (a) monometallic, (b) bimetallic, and (c) trimetallic nanoparticle datasets. "F" represents feature.

References

[1] Christian Thurau, Kristian Kersting, Mirwaes Wahabzada, and Christian Bauckhage. Convex non-negative matrix factorization for massive datasets. *Knowledge and Information Systems*, 29:457–478, 2010.

[2] Sebastian Mair, Ahcène Boubekki, and Ulf Brefeld. Frame-based data factorizations. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2305–2313. PMLR, 2017.

[3] Sebastian Mair and Ulf Brefeld. Coresets for archetypal analysis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch{é} Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1–9. Curran Associates, Inc., 2019.

[4] Ruijian Han, Braxton Osting, Dong Wang, and Yiming Xu. Probabilistic methods for approximate archetypal analysis. *Information and Inference: A Journal of the IMA*, 12:466–493, 2023.

[5] Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012.

[6] Sebastian Mair and Jens Sjölund. Archetypal analysis++: Rethinking the initialization strategy. *Transactions on Machine Learning Research*, pages 1–27, 2024.

[7] Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10:180–184, 5 1985.

[8] Dan Feldman. *Core-Sets: Updated Survey*, pages 23–44. Springer, Cham, 2020.

[9] Vinayak Abrol and Pulkit Sharma. A geometric approach to archetypal analysis via sparse projections. In *Proceedings of the 37th International Conference on Machine Learning*, pages 42–51. PMLR, 2020.

[10] J. G. Kreer. A question of terminology. *IRE Transactions on Information Theory*, 3:208, 1957.

[11] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[12] F. W. Scholz and M. A. Stephens. K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, 82:918–924, 2012.

[13] W J Conover. Several k-sample Kolmogorov-Smirnov tests. *The Annals of Mathematical Statistics*, 36:1019–1026, 1965.

[14] T. W. Anderson. On the distribution of the two-sample Cramer-von Mises criterion. *Annals of Mathematical Statistics*, 33:1148–1159, 1962.

[15] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6:366–422, 1960.

[16] Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5:445–463, 1977.

[17] Bent Fuglede and Flemming Topsøe. Jensen-shannon divergence and hubert space embedding. In *IEEE International Symposium on Information Theory - Proceedings*, page 31, 2004.

[18] Amanda Barnard and George Opletal. Gold nanoparticle data set. v1., 2019.

[19] Amanda Barnard and George Opletal. Palladium nanoparticle data set. v1., 2019.

[20] Amanda Barnard, Baichuan Sun, and George Opletal. Platinum nanoparticle data set. v2., 2019.

[21] Jonathan Ting, Amanda Barnard, and George Opletal. AuPd nanoparticle data set. v1., 2023.

[22] Jonathan Ting, Amanda Barnard, and George Opletal. AuPt nanoparticle data set. v1., 2023.

[23] Jonathan Ting, Amanda Barnard, and George Opletal. PdAu nanoparticle data set. v1., 2023.

[24] Jonathan Ting, Amanda Barnard, and George Opletal. PdPt nanoparticle data set. v1., 2023.

[25] Jonathan Ting, Amanda Barnard, and George Opletal. PtAu nanoparticle data set. v1., 2023.

[26] Jonathan Ting, Amanda Barnard, and George Opletal. PtPd nanoparticle data set. v1., 2023.

[27] Kaihan Lu, Jonathan Ting, Amanda Barnard, and George Opletal. AuPdPt nanoparticle data set. v1., 2023.

[28] Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 172–177, 2010.