Electronic Supplementary Information

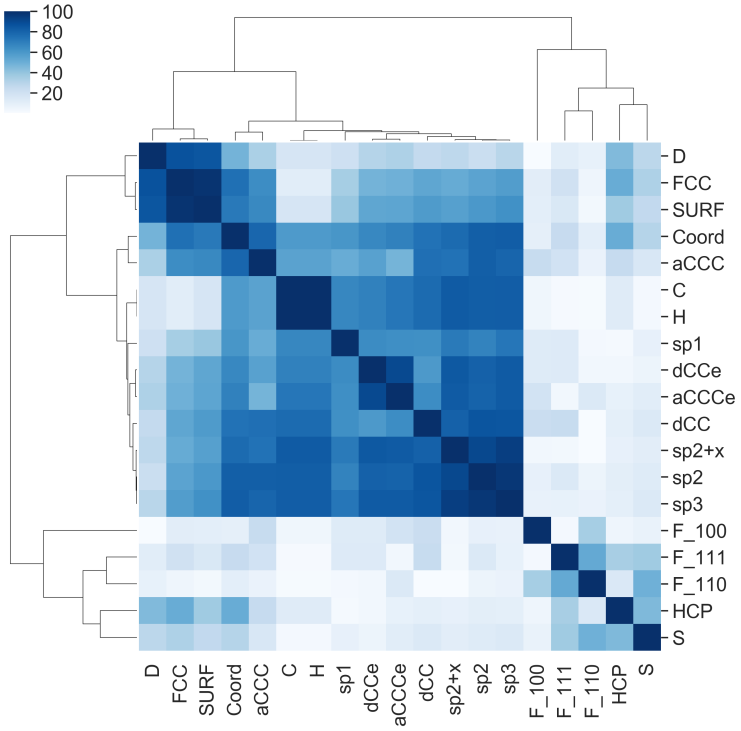
Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbor embedding and machine learning

Amanda S. Barnard* and George Opletal

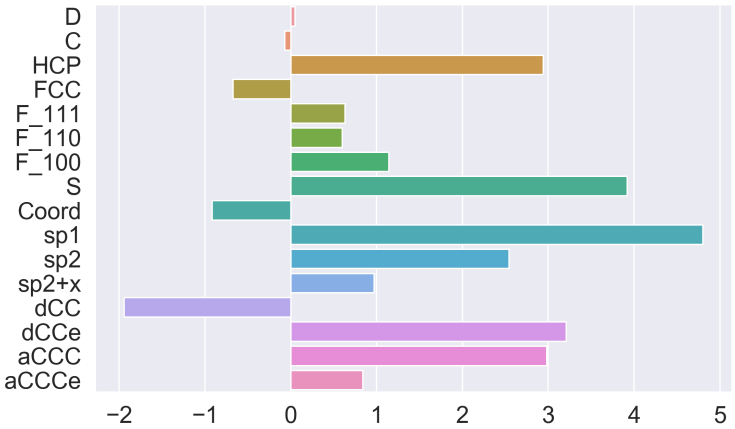
CSIRO Data61, Docklands, Victoria 3008, Australia

E-mail: amanda.barnard@data61.csiro.au

Contained herein are the correlation matrix (Figure S1(a)) used to eliminate reductant features, and the skew map showing the asymmetry of the distribution of the remaining features (Figure S1(b)), as mentioned in the main text. This is followed by illustrations of the geometric shapes and their categorial labels (Figure S2) and the t-SNE scatter plot encoded with the categorial shape labels (Figure S3). The remaining plots are the t-SNE maps encoded with the absolute value of the remaining features not shown in the main text (Figures S4, S5 and S6), and the results for t-SNE maps using different perplexities (Figure S7) and different data preparation protocols (Figure S8) such as including the redundant features, non-normalisation and initialisation with principle components analysis (Figure S9), and variations of the early exaggeration parameter (Figures S10 and S11). Hyper-parameters for the regressors used in the main text are provided in Table S1. The distribution of features are provided in Figures S12, S13 and S14. Results for previous generation models identified by TPOT (as described in the main text) are provided in Figure S15.



(a)



(b)

Figure S1: (a) Correlation matrix identifying the correlated features of the nanodiamond dataset; (b) Skew map showing the maximum variance in the un-correlated features of the nanodiamond dataset.

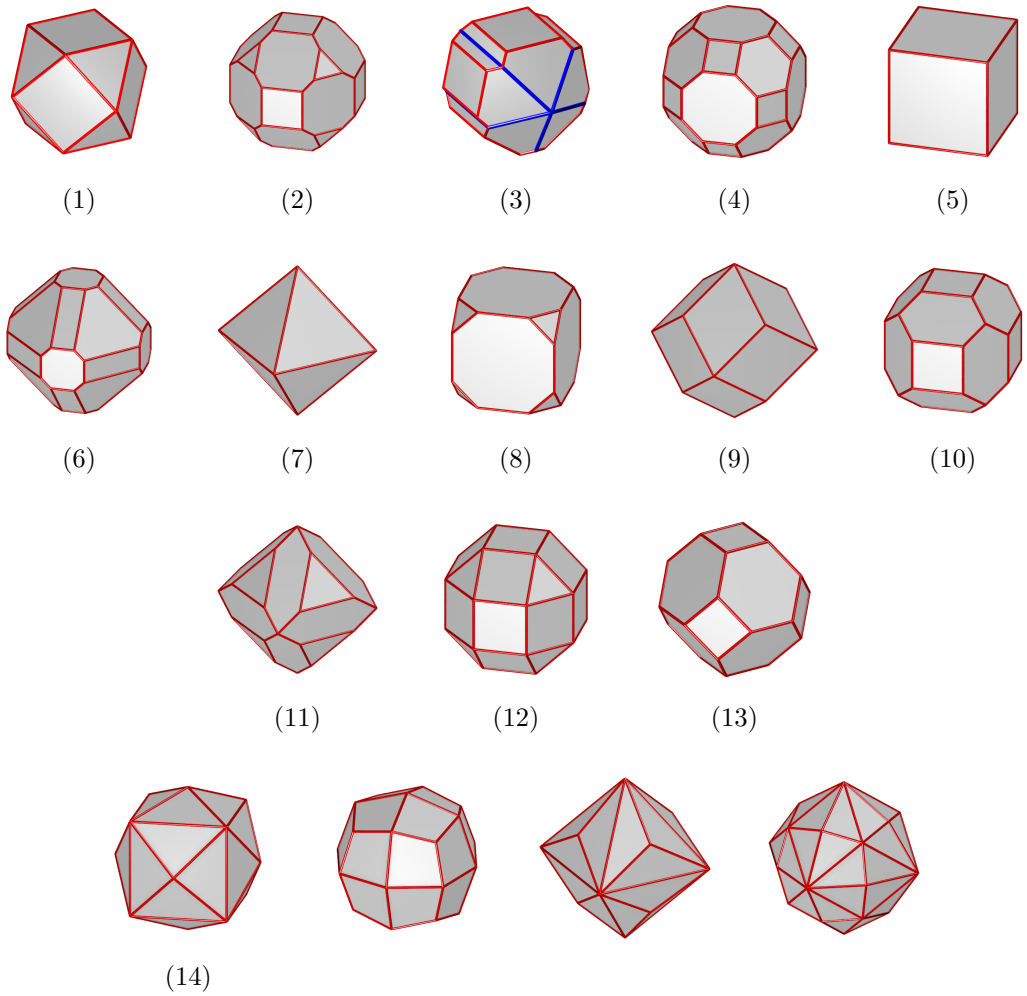
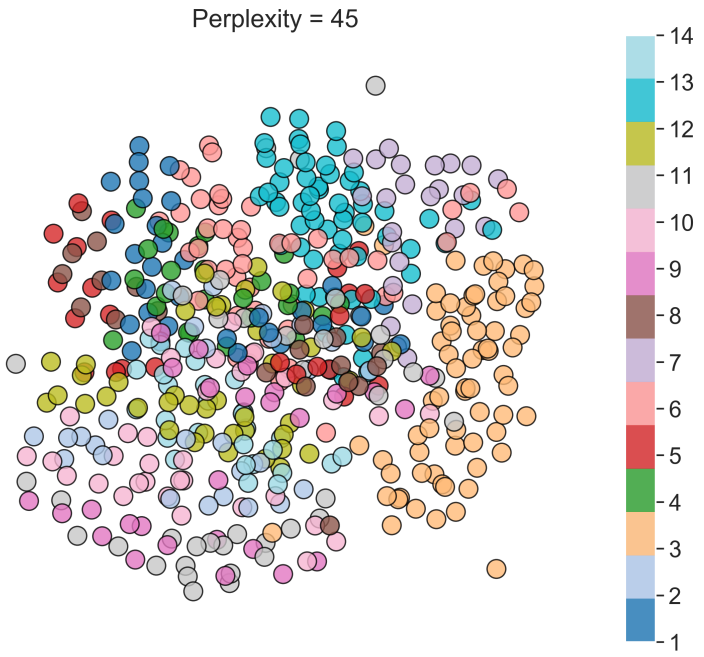
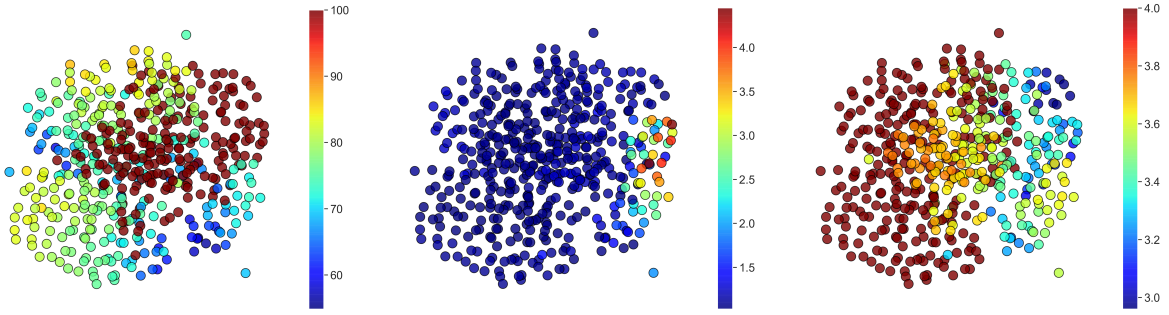


Figure S2: Shapes included in the nanodiamond data set, including the (1) cuboctahedron, (2) modified-truncated rhombic dodecahedron, (3) decahedron (with hcp twin planes highlighted in blue), (4) great rhombicuboctahedron, (5) cube, (6) modified-truncated octahedron, (7) octahedron, (8) truncated cube, (9) rhombic dodecahedron, (10) rhombi-truncated cube, (11) rhombi-truncated octahedron, (12) small rhombicuboctahedron, (13) truncated octahedron and (14) high index shapes including the tetrakis hexahedron, the trapezohedron, the trisoctahedron and the hexakis octahedron.



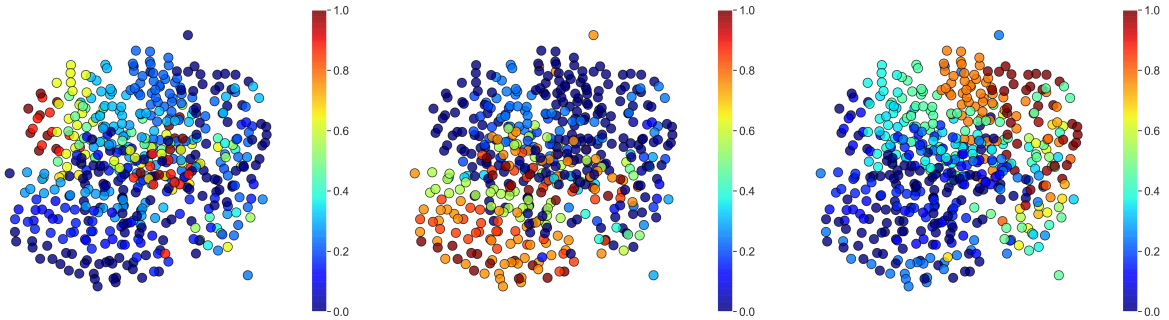
(a)

Figure S3: t-SNE map of the 16-dimensional nanodiamond data set encoded with the shape, as defined in Figure S2. Note that the shape is an external categorial label that was not used to train the model.



(a) (b) (c)

Figure S4: t-SNE map of the 16-dimensional nanodiamond data set encoded with absolute values of structural features including (a) the concentration of C atom, (b) the sphericity (anisotropy), and (c) the average coordination number of C atoms.



(a) (b) (c)

Figure S5: t-SNE map of the 16-dimensional nanodiamond data set encoded with absolute values of structural features including (a) the fraction of (100) surface area, (b) the fraction of (110) surface area, and (c) the fraction of (111) surface area.

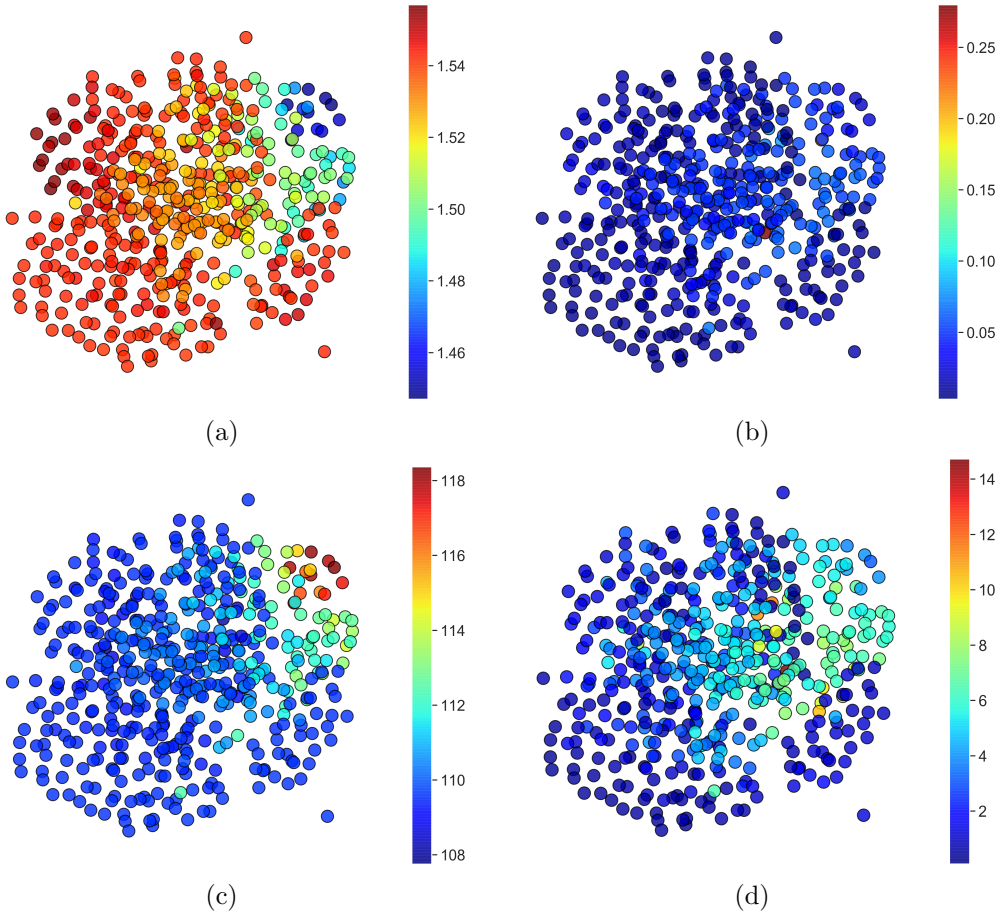


Figure S6: t-SNE map of the 16-dimensional nanodiamond data set encoded with absolute values of structural features including (a) the C–C bond length, (b) the statistical error in the C–C bond length, (c) the C–C–C bond angle, and (d) the statistical error in the C–C–C bond angle.

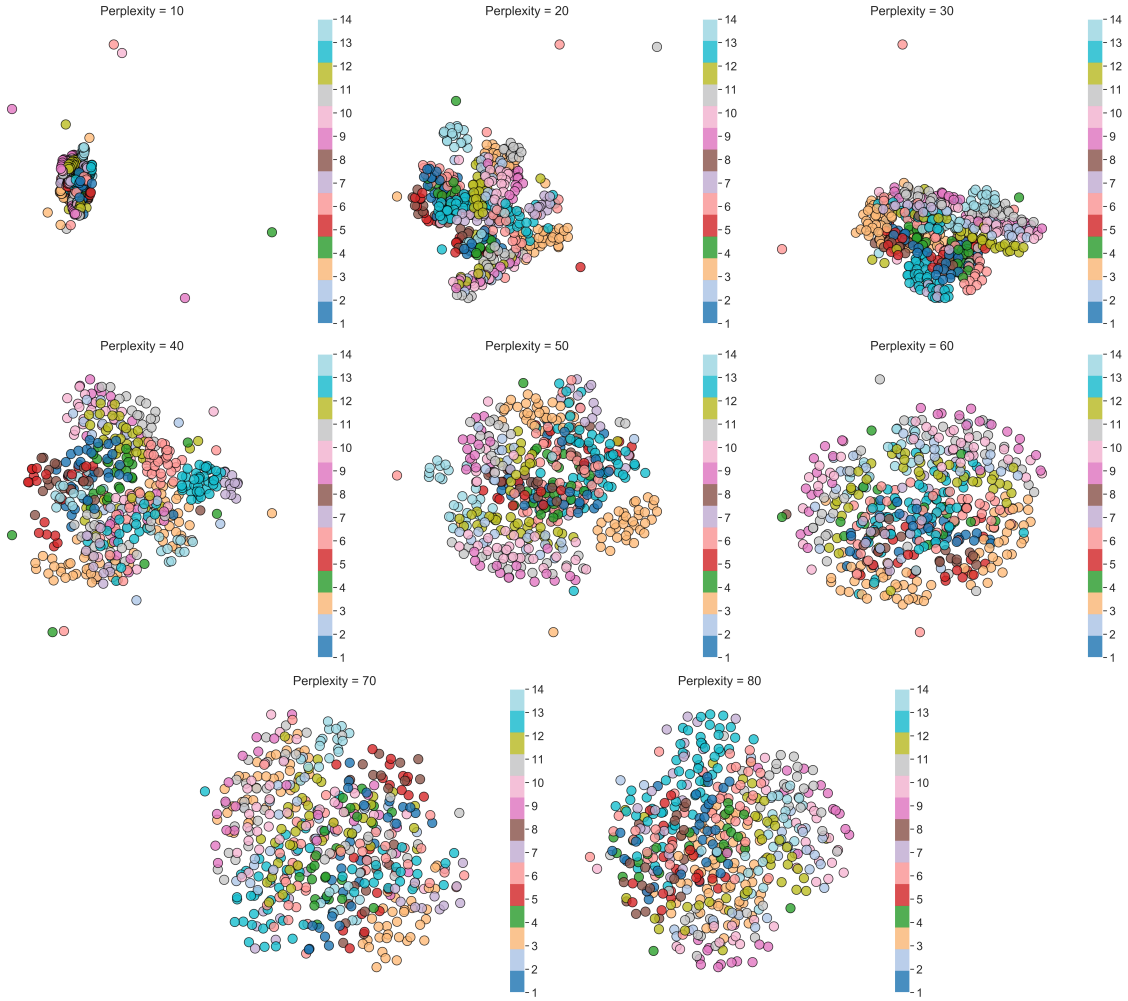
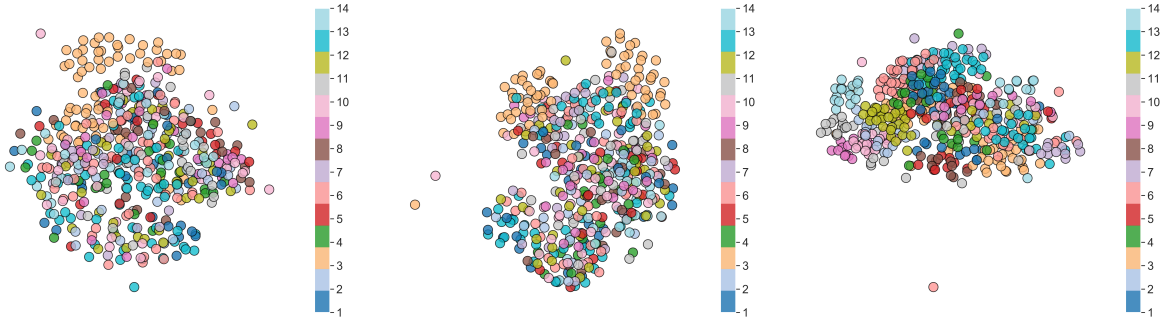
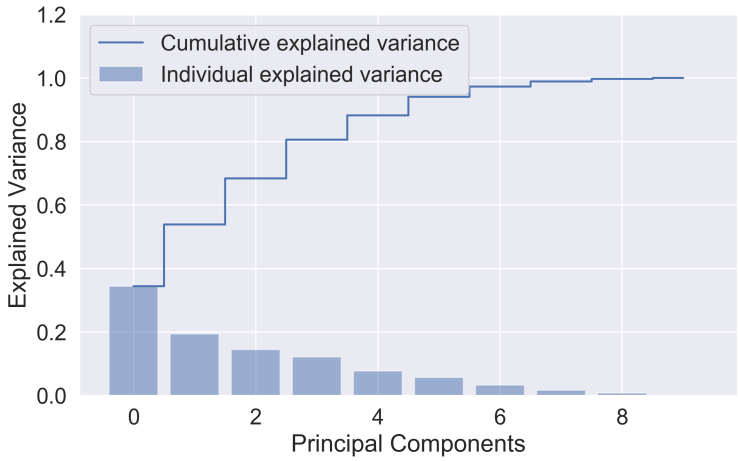


Figure S7: t-SNE map of the 16-dimensional nanodiamond data set, encoded with the shape (external label), as defined in Figure S2, trained with different perplexities.



(a) (b) (c)

Figure S8: t-SNE map of the 16-dimensional nanodiamond data set, encoded with the shape (external label), as defined in Figure S2, trained with a perplexity of 45 and (a) strongly correlated features included, (b) un-normalisation of the features, and (c) initialisation with PCA using 8 principle components instead of features.



(a)

Figure S9: Cumulative explained variance of the principle components of the nanodiamond data set calculated with principle component analysis.

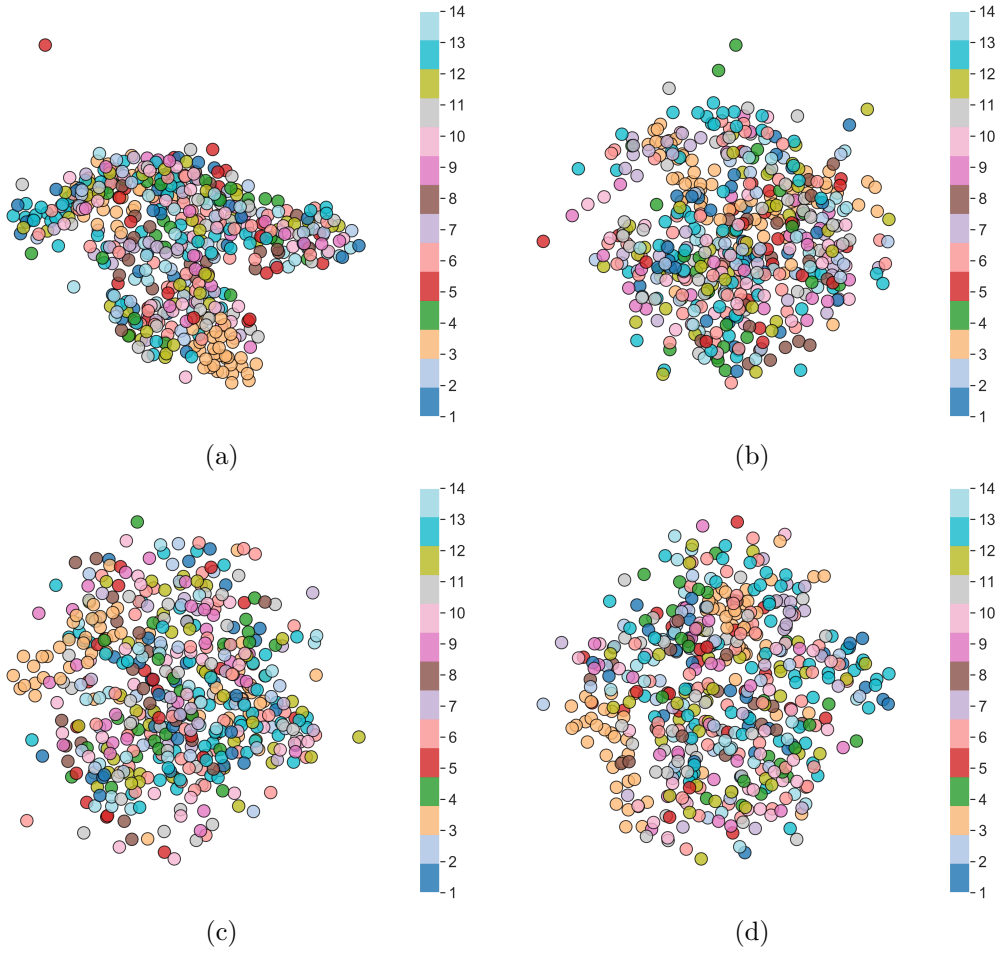
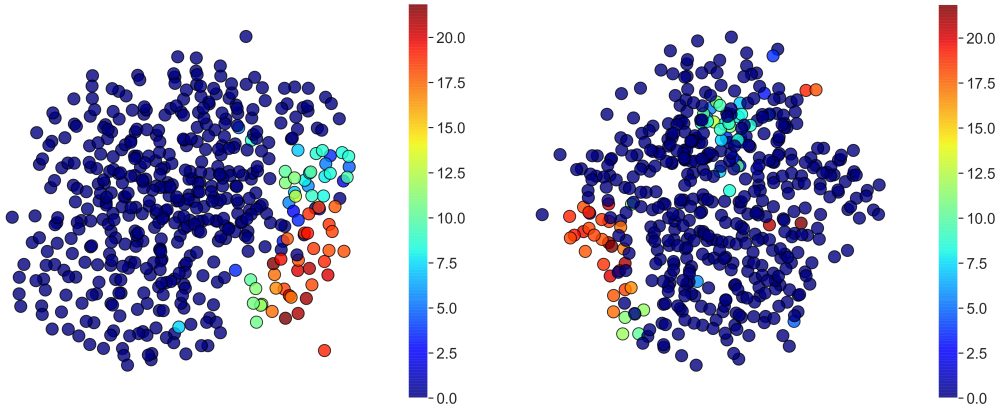


Figure S10: t-SNE map of the 16-dimensional nanodiamond data set using a perplexity of 25, encoded with absolute values of structural features including (a) an early exaggeration parameter of 12 (default), (b) an early exaggeration parameter of 15, (c) an early exaggeration parameter of 20, and (d) an early exaggeration parameter of 25. Note that while this approach fails to capture the groupings in the distribution that can be drawn out by changing the perplexity.



(a) (b)

Figure S11: t-SNE map of the 16-dimensional nanodiamond data set using (a) a perplexity of 45 with an early exaggeration parameter of 12 (default), and (b) a perplexity of 25 with an early exaggeration parameter of 25 to artificially disperse the distribution, encoded with absolute values the HCP concentration. It is clear from this comparison that dispersion using the perplexity retains the intrinsic similarity of the data in the feature space, as the perplexity is designed to balance attention between local and global characteristics of the data set, while dispersion using the early exaggeration parameter is not, and results in fragmentation of a group of like-samples.

Table S1: Hyper-parameters used for the regressors discussed in the main text

Hyper-parameter	Gradient Boosting	Extra Trees	Gradient Boosting
α	0.95	N/ A	0.75
Learning rate	0.1	N/A	0.1
Maximum depth	5	None	8
Maximum features	0.8	0.65	0.3
Minimum sample leaves	4	1	4
Minimum sample splits	18	4	9
Number of estimators	100	100	100
Sub-sample	0.4	N/A	0.6
Bootstrapping	False	True	False

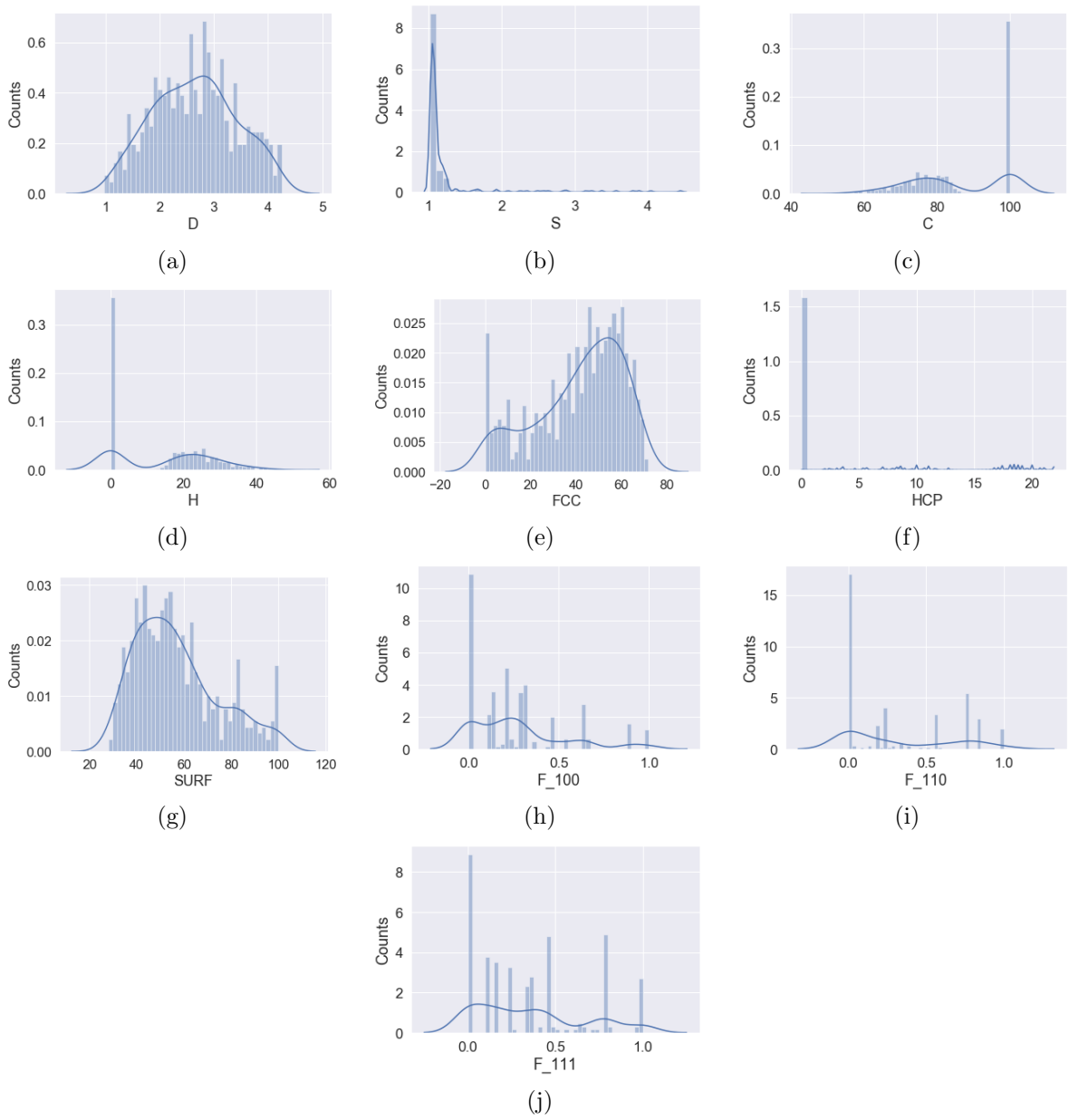


Figure S12: The distributor of the absolute values of structural features including (a) the diameter in nm, D, (b) the sphericity, S, (c) the fraction of carbon atoms C, (d) the fraction of hydrogen atoms, H, (e) the fraction of face-centered cubic atoms, FCC, (f) the fraction of hexagonal close-packed atom, HCP, (g) the fraction of surface atoms, SURF, (h) the fraction of {100} surface area, F_100 , (i) the fraction of {110} surface area, F_110, and (j) the fraction of {111} surface area, F_111.

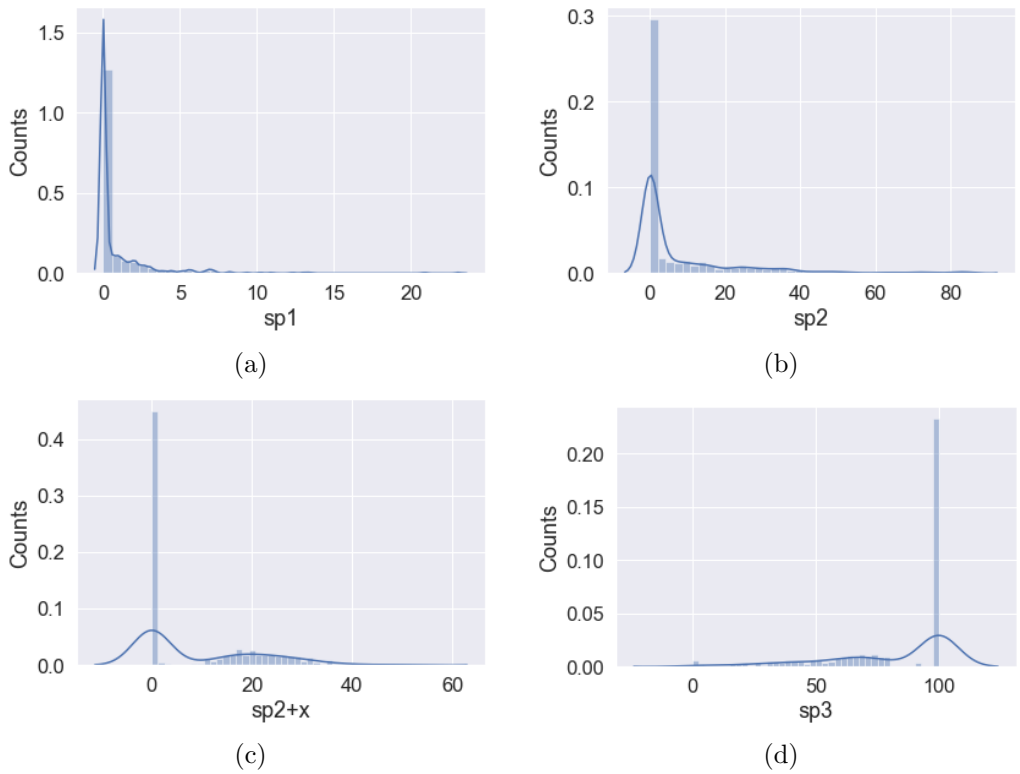


Figure S13: The distributor of the absolute values of chemical features including (a) the concentration of sp^1-hybridized atoms, sp1, (b) the concentration of sp^2-hybridized atoms, sp2, (c) the concentration of sp^{2+x}-hybridized atoms, sp2+x, and (d) the concentration of sp^3-hybridized atoms, sp3.

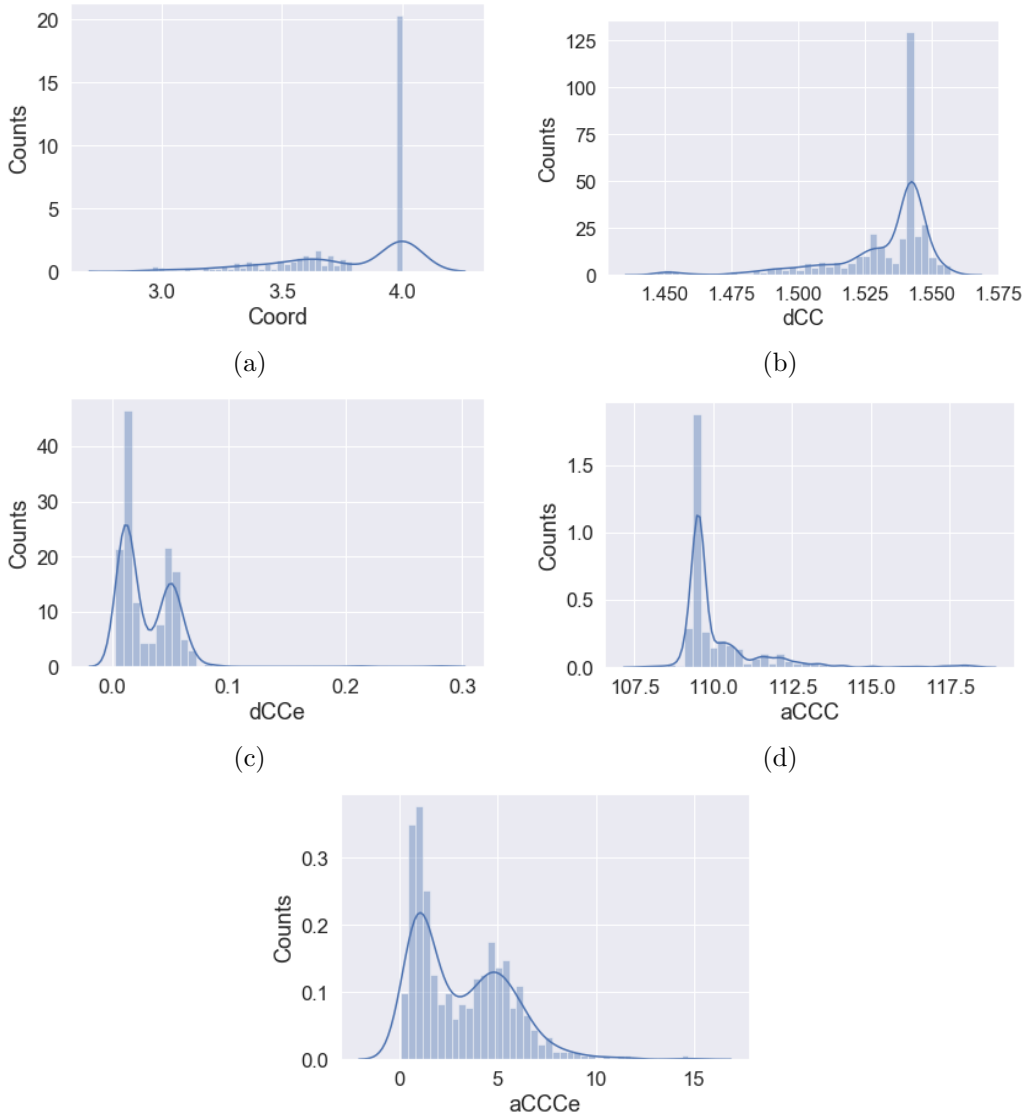


Figure S14: The distributor of the absolute values of statistical features including (a) the average C–C coordination number, Coord, (b) the average C–C bond length in Å, dCC, (c) the statistical error in the average C–C bond length in Å, dCCe, (d) the average C–C–C bond angle in Degrees, aCCC, (c) the statistical error in the average C–C–C bond length in Degrees, aCCCe,

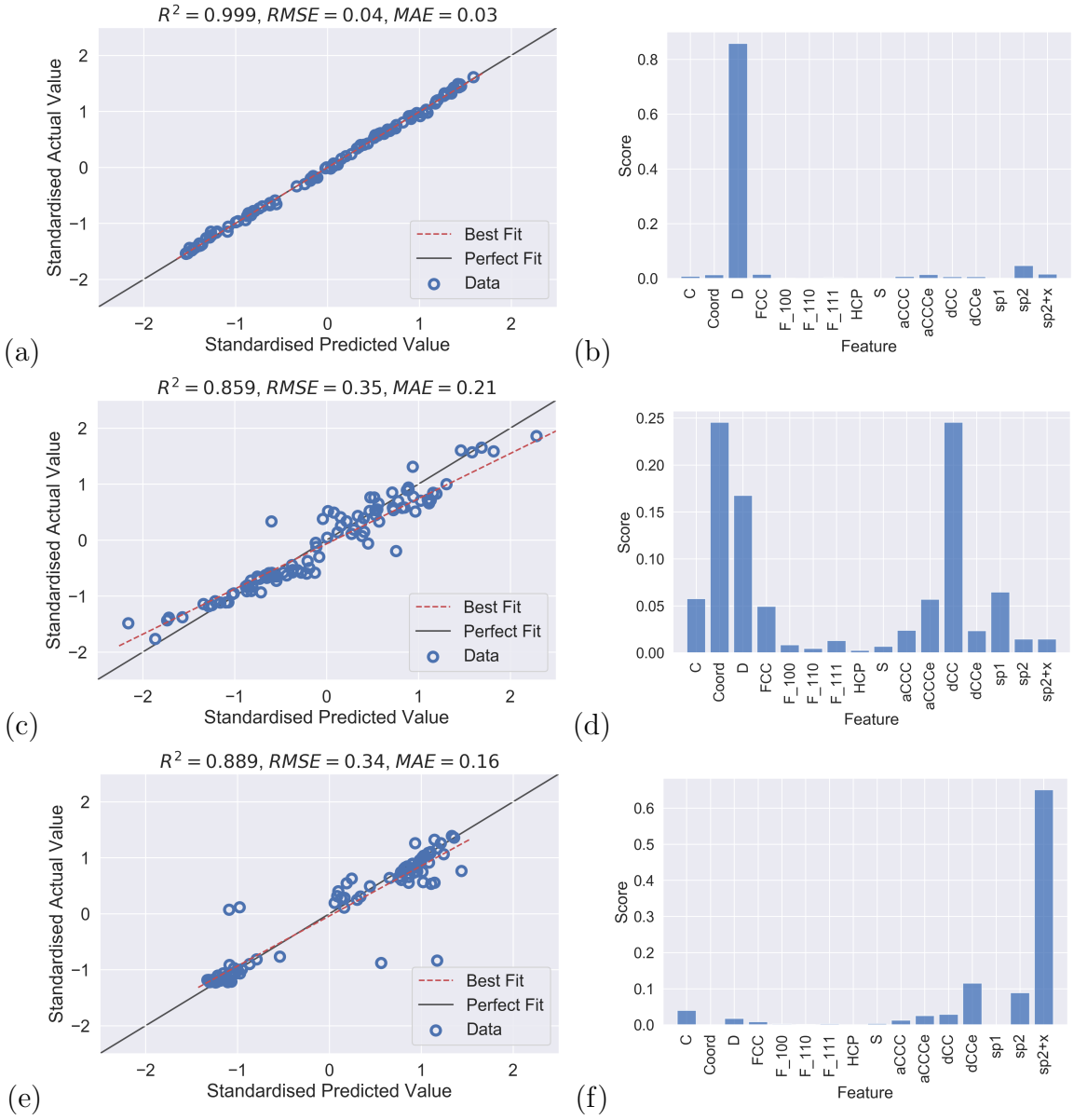


Figure S15: Results for the previous generation model fits identified using TPOT, for the testing data from the 80/20 split (a, c, e) and feature importance histograms (b, d, f) for the probability (a,b) ionisation potential (c,d) and band hap (e,f), respectively. The alternative models are described in the main text.