MRS Communications (2019), *9*, 730–736
© Materials Research Society, 2019
doi:10.1557/mrc.2019.36

MRS Communications

Research Letter

Identifying hidden high-dimensional structure/property relationships using self-organizing maps

Amanda S. Barnard, **Benyamin Motevalli,** and **Baichuan Sun,** CSIRO Data61, Door 34 Village Street, Docklands, VIC 3008, Australia
Address all correspondence to Amanda S. Barnard at **amanda.barnard@data61.csiro.au**

(Received 3 January 2019; accepted 15 March 2019)

Abstract

Unlike other data intensive domains, understanding distributions, trends, correlations, and relationships in materials data sets typically involves navigating high-dimensional spaces with only a limited number of observations. Under these conditions extracting structure/property relationships is not straightforward and considerable attention must be given to the reduction of feature space before predictions can be made. Here we have used Kohonen networks (self-organizing maps) to identify hidden structure/property relationships in computational sets of twinned and single-crystal diamond nanoparticles based on structural similarity in multiple dimensions, and confirmed the importance of a limited number of surface chemical features using regression.

Introduction

Materials data sets are inherently complicated and multi-dimensional, so it is challenging to discover and exploit the hidden non-linear correlations[1–8] that form the basis for useful structure/property relationships. Dealing with a multi-dimensional data sets comes with unique problems, which are often exacerbated in this domain due to most data sets being limited to a small number of observations or simulations; typically only hundreds to thousands.

Numerous quantitative features can be extracted using various analytical and statistical techniques available today, making it relatively straightforward to compile a comprehensive list of descriptors during computational or experimental studies that drawn on multiple characterization tools. However, just because a feature is easy to quantify does not mean that it is an important driver of structure/property relationships; nor does its inclusion guarantee a better understanding of the material or its function. In most cases it is desirable to base structure/property relationship on as fewer structural features as possible, to make it easier to guide the development of bespoke materials targeted to specific applications.

Feature engineering is a powerful way of reducing the number of dimensions by eliminating materials features that are strongly correlated, or by replacing groups of features with a new one, but this requires significant domain knowledge and may not be possible in all cases. Not all features/dimensions can be reduced in this way, and this approach risks introducing additional biases since these decisions are subjective and researchers suffer from confidence bias more than the general population.

An alternative is to use an appropriate machine learning algorithm to reduce high-dimensional materials data to a more manageable low-dimensional set, and to facilitate a more straightforward comparison of structural features with property labels. One example is Kohonen network,[9] or self-organization map (SOM), which is an unsupervised artificial neural network[10] for non-linearly mapping high-dimensional spaces into low-dimensional spaces with the advantage of retaining the intrinsic topological relationship of the input data set. SOMs are ideal for visualizing multi-dimensional data sets in a single two-dimensional image,[11] and have been recently used to create surface texture maps of nanoparticles for uses as materials fingerprints.[12]

In the current work we demonstrate the reduction of materials data using feature engineering and SOMs to uncover hidden structure/property relationships in computational sets of twinned and single-crystal diamond nanoparticles, based on structural similarity in multiple dimensions. Our objective is to identify which structural feature is most promising for tuning charge transfer properties including the ionization potential, electron affinity, electronic band gap, and electronegativity. As we will see, many of the characteristics previously considered to be important do not correlate strongly with these properties, and so synthesis and processing strategies will need to be revised.

Self-organizing maps

The basic units of an SOM are neurons which are best organized on hexagonal grids. All neurons can be arranged in a planar (approximate) rectangle, with periodic boundary conditions to connect the upper and lower boundaries, and the right and left boundaries, so that the SOM plane occupies the surface of a torus. The weights of all neurons are initialized using random numbers. During each training step every neuron

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

competes against all others until each original data point finds the one neuron that is closest to it in Euclidean space, referred to as the best matching unit (BMU). Given an input data point x, and the weight of neuron i, j is w^{ij}, then the Euclidean distance D is:

$$D = \sqrt{\sum_{v=1}^{d} (x_v - w_v^{ij})^2}$$

where v is each component of the vector and D is the dimension of the normalized data set. Once the BMU is located, the weights of all neurons centered on it are updated. Initially all neurons on the SOM are updated, until only the BMU is updated at the conclusion of training. This ensures the radius of neighborhood $\delta(t)$ decreases with each subsequent iteration step, t. In addition to this a linear relation can be adopted with a pre-set maximum number of epochs, n_{epoch}, to train the SOM, such that:

$$\frac{\delta(t) - \delta_0}{t - 1} = \frac{\delta_{n_{\text{epoch}}} - \delta_0}{n_{\text{epoch}} - 1}$$

where δ_0 is the initial radius which has been set to half size of the SOM, $t = 2, 3, 4, \ldots, n_{\text{epoch}}$, and $\delta_{n_{\text{epoch}}}$ is determined by the boundary condition such that it equals one at the last step of training. In addition to decay of the neighborhood radius, the learning rate $L(t)$ for updating weights on each neuron also decreases with each iteration of t. Weights are updated faster early in the training, and slower toward the end of training, based on:

$$\frac{L(t) - L}{t - 1} = \frac{L_{n_{\text{epoch}}} - L_0}{n_{\text{epoch}} - 1}$$

where L_0 is the initial learning rate, $L_{n_{\text{epoch}}}$, is the constant and chosen to ensure the training is very fine grained. As a result, the updating procedure is given by

$$w^{ij}(t+1) = w^{ij}(t) + L(t) \exp\left[\frac{-D_x(t)^2}{2\delta^2(t)}\right][x - w^{ij}(t)]$$

where $D_x(t)$ is the Euclidean distance of a neuron at i, j from the BMU of x at step t.

The total number of epochs, n_{epoch}, is an important hyperparameter of SOM training. A large number of epochs will result in over-training, leading to a waste of computational resources, but a small number of epochs will result in under-training, and dissimilar data points may be adjacent on the final SOM reducing the resolution. It is possible to measure the number of void neurons which have no weight, and stop the training when this number stops decreasing over a threshold after, for instance, $n_{\text{epoch}} = 5$, to improve efficiency and autonomy.[13]

Feature analytics

To demonstrate the process of dimension reduction with SOMs we have combined two previously prepared computational data sets that have been made publicly available.[14,15] We have selected 355 unpassivated/reconstructed and hydrogen passivated diamond nanoparticles between 1.5 and 4.5 nm in size (spanning the experimentally observed size range), and 13 different shapes enclosed by different combinations of low index (100), (110), and (111) facets; with and without internal twinning. These sets are characterized by 32 structural features, five numerical property labels (the energy that is used to calculate the probability of observation, the ionization potential, the electron affinity, the band gap, and the electronegativity), and one external categorical label (the nanoparticle shape). This gives us a 38 dimensional data set. The method used originally to generate this particular data set is described in detail elsewhere.[16] The next step is to clean the data and condense the feature space to reduce the risk of overfitting. Overfitting occurs when a model learns the detail and noise in the training data so well that it negatively impacts the performance of the model on new data, making it effectively useless.

To begin we clean the data by removing outliers and excluding structural features that contain missing entries, as they are not defined for all of the data points (NaN). For example, the bond length of sp^2 hybridized carbon atoms is not defined for a fully passivated nanodiamond where all atoms are no sp^2 hybridized atoms. We also exclude secondary features associated with the uncertainties of primary features (e.g., where the bond angle $\Theta_{C-C-C} = 109.13 \pm 3.56°$, we retain only the feature listing 109.13°). This process reduced the number of dimensions to six labels and 14 features, including the nanoparticle diameter (D), number of carbon (N_{Carbon}), and hydrogen (N_{Hydrogen}) atoms, the fraction of (100), (110), or (111) surface area (F_{100}, F_{110}, and F_{111}), the sphericity (S), the average C–C coordination number (Coord), the fraction of sp^1, sp^2, sp^{2+x}, and sp^3 hybridized atoms, the average C–C bond length (d_{C-C}), and the average C–C–C bond angle (Θ_{C-C-C}).

We then normalized these features and calculated the correlation matrix (Fig. 1). This confirmed that D and N_{Carbon} are strongly correlated, the fraction of sp^2 atoms, d_{C-C} and Θ_{C-C-C} are strongly correlated, and Coord and the fraction of sp^3 atoms are strongly correlated. We therefore exclude N_{Carbon}, d_{C-C}, Θ_{C-C-C}, and the fraction of sp^3 (which are bulk-like); retaining D, sp^2 fraction, and Coord (which are surface-like). This reduces the number of structural dimensions to 10, but it is interesting to note that the fraction of (111) surface area is not *strongly* correlated with the fraction of sp^2 atoms, as would be typically expected based on previously published studies.[17,18]

Dimension reduction

We then used these remaining structural features to train SOMs on 500 × 500 grids over five epochs; reducing the dimensionality to 2, and facilitating the visualization of the results in a

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

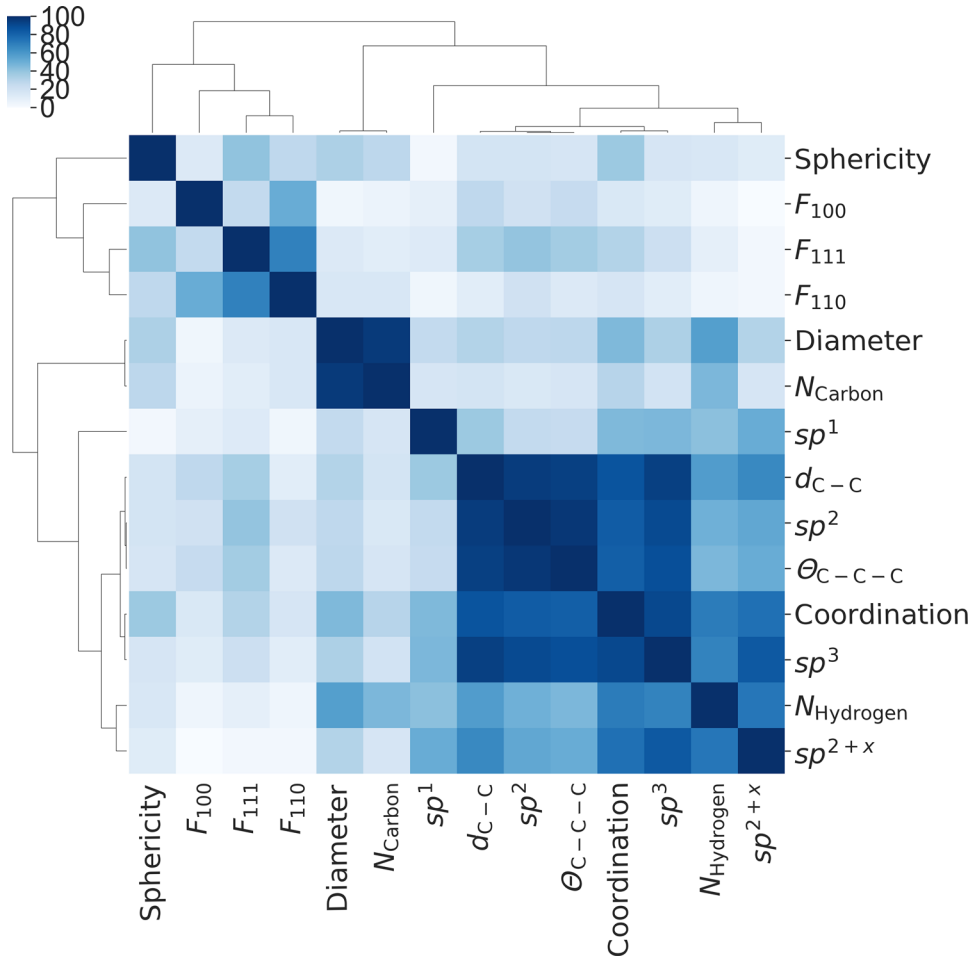


Figure 1. Hierarchical cluster map of the materials' features describing the diamond nanoparticles after the removal of outliers, NaNs, and secondary features.

scatter plot. The grid size is a hyper-parameter that determines the available number of possible neurons for matching. Too few and data are poorly mapped, and too many increases the computational cost. Given that every *nanoparticle* (or row/observation) gets mapped to the winning neuron (or BMUs) in the SOM and its influence spreads to neighboring neurons, the grid size is typically approximately commensurate with number of samples (nanoparticles). We tested values around this size and found that a grid size of 500×500 reduced the amount of overlap of the data points (giving them a better spread of possible BMUs) and aided in visualization and interpretation, without significantly increasing the computational cost.

The final distribution of the points reflects the degree of similarity of each structure in 10-dimensional structure space; adjacent points are very similar, and separated points are very different. To appreciate how some important structural features are mapped on to this manifold we can encode each point with a color representing the actual magnitude of the structural feature prior to normalization. This has been done for four important features: the diameter (in nm), the number of surface hydrogen atoms, the fraction of sp^2 atoms, and the fraction of sp^{2+x} atoms (Fig. 2). All of these features describe surface effects, including the diameter which is a convenient measure of size related to the surface-to-volume ratio.

From Fig. 2 we can see that the large nanoparticles do not tend to be mapped together, which is counterintuitive for a community that often thinks in terms of size-dependent structure/property relationships. The fraction of sp^2 and sp^{2+x} atoms are however mapped to the same region of the space (recalling that SOMs have periodic boundary conditions), and exhibit similar relationships, though the distribution is not reflected by the size or the surface hydrogenation. More SOM feature distributions are provided for comparison in the supplemental information, where we can see that the features displayed in Fig. 2 are distributed very differently to the fraction of different surface facets, the local bonding arrangements, and the overall sphericity or under-coordination of the particles.

In addition to visualizing the feature space we can also encode the SOMs by the target labels (the properties, in our

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

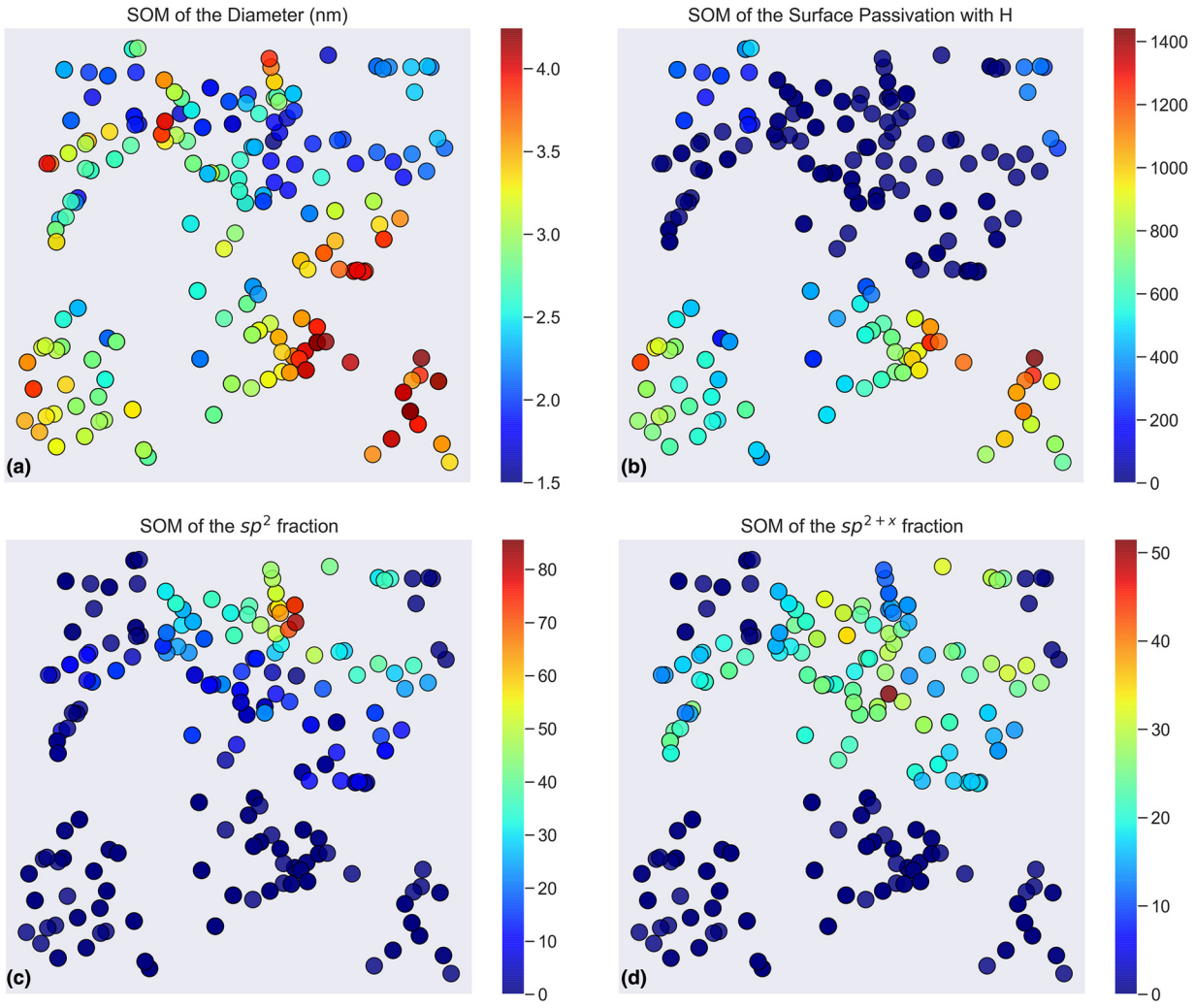


Figure 2. SOMs of the 10-dimensional diamond nanoparticle data set after the removal of strongly correlated features, encoded by the absolute value of the diameter (a), the number of surface hydrogen atoms (b), fraction of sp^2 hybridized atoms (c), and fraction of sp^{2+x} hybridized atoms (d).

structure/property relationships), specifically the four charge transfer properties (Fig. 3). Here we can see that the electron affinity, electronic band gap, and the electronegativity present very similarly. The distribution of the electronic band gap and the electronegativity in 10-dimensions are almost identical, with the maps differing only by the scale of the encoded properties. When we compare these property labels (Fig. 2) to the features highlighted in Fig. 1 we can see that the electron affinity, electronic band gap, and the electronegativity are clearly related to the fraction of sp^{2+x} hybridized atoms, not to the size or sp^2 content as is usually assumed. Comparing with other feature (structure) maps and label (property) maps contained in the supplemental information also confirms that these properties are not strongly associated with the fraction of different types of surface facets or the overall nanoparticle shape, which has also been previously assumed. The probability of observation (provided in the supplemental information) is however, clearly related to the diameter, which is consistent with the fact that larger nanoparticles are more stable.

Feature importance

To confirm the relationship between the fraction of sp^{2+x} hybridized atoms and the charge transfer properties we undertook regression using a random forest (RF) regressor, which is a meta estimator that fits a number of classifying decision trees on sub-samples of the data set and averages them to improve accuracy and avoid over-fitting.[19] We have used 100 estimators with a minimum of two split samples, and bootstrapping to randomly select features to build trees and reduce problems with feature multi-collinearity. A fivefold cross validation was used to confirm the accuracy and generality of the model, with 80% of the data used for training and 20% for testing. While other regression algorithms are available and could also be applied, this regressor is particularly well suited to our problem

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

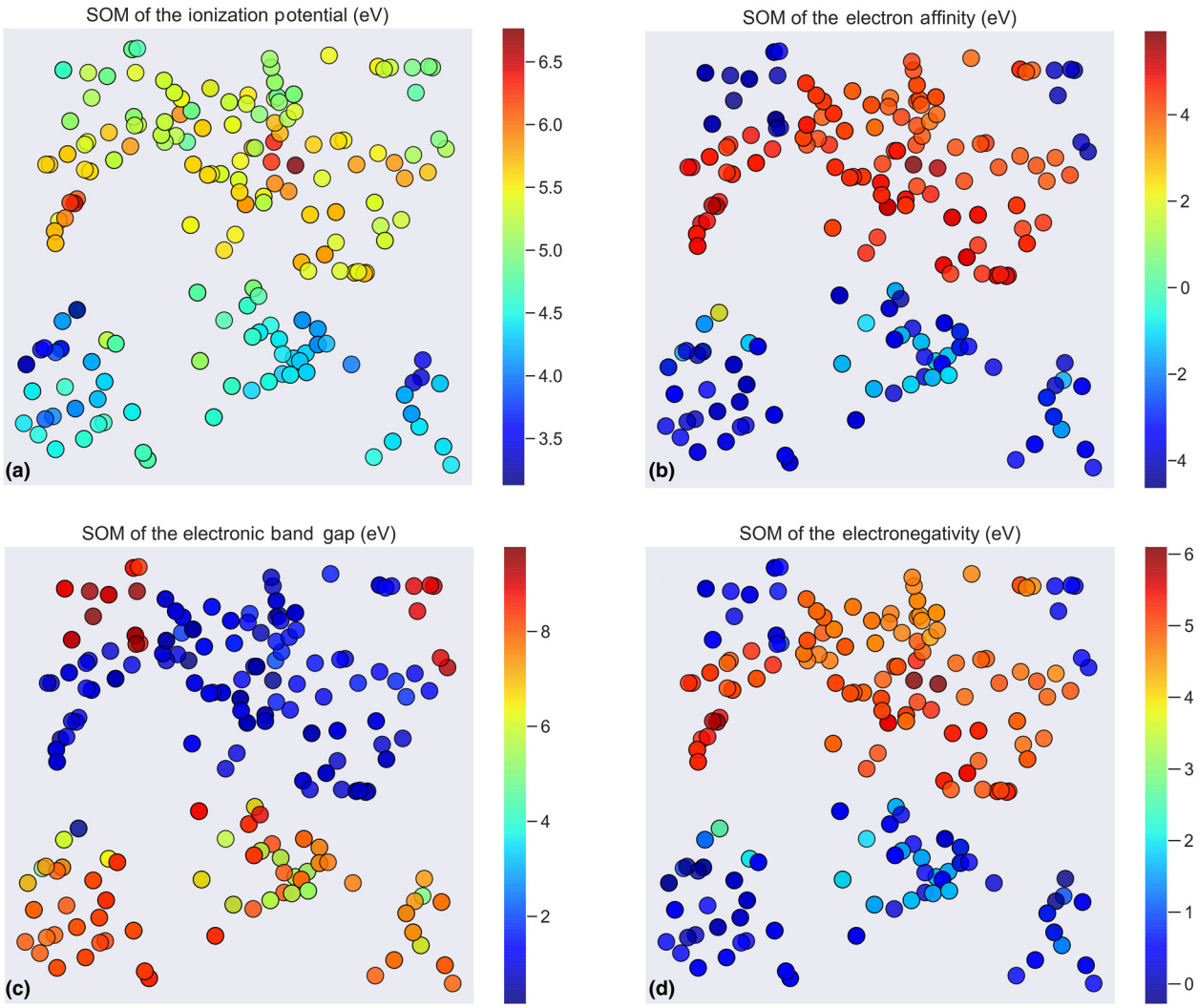


Figure 3. SOMs of the 10-dimensional diamond nanoparticle data set after the removal of strongly correlated features, encoded by the absolute value of the energy of the ionization potential (a), electron affinity (b), electronic band gap (c), and electronegativity (d).

since the data set has a high variance and we know from the input data set that we have different types of nanoparticles (twinned/un-twinned, passivated/reconstructed, etc.) that will likely all coexist in real diamond nanoparticle powders.

The RF regressor achieved excellent results for the testing (training) sets, as measured using the R^2 scores. For the ionization potential we obtained testing (training) values of $R^2 =$ 0.906 (0.899); for the electron affinity we obtained $R^2 =$ 0.963 (0.947); for the band gap we obtained $R^2 = 0.948$ (0.933); and for the electronegativity $R^2 = 0.966$ (0.955). When we examine the importance of each feature used in the RFs we find that the regression confirms the structure/property relationships uncovered by the SOMs. Figure 4 shows the feature importance histograms for the RF of each charge transfer property.

The feature importance histograms provide three important findings. First, in agreement with the SOMs, the fraction of sp^{2+x} hybridized atoms is an important feature in controlling all of the charge transfer properties, with the exception of the ionization potential (again, in agreement with the SOMs). Second, that we can confidently reduce the number of dimensions to three or less, since the number of hydrogen atoms, the fraction of sp^2 hybridized atoms and sp^{2+x} hybridized atoms dominate the models and will be sufficient to form the basis of structure/property relationships. These are all surface features. While the importance of sp^{2+x} hybridized atoms has been largely overlooked in the literature to date (which has tended to focus on sp^2 hybridized atoms or surface functionalization), the importance of these three materials features is generally consistent with over a decade of domain knowledge,[20,21] suggesting that the ability to either passivate the surfaces or moderate the type and extent of surface reconstructions is paramount. Finally, these machine-learning results also highlight that the *Diameter* of the nanoparticle is not an important driver of

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

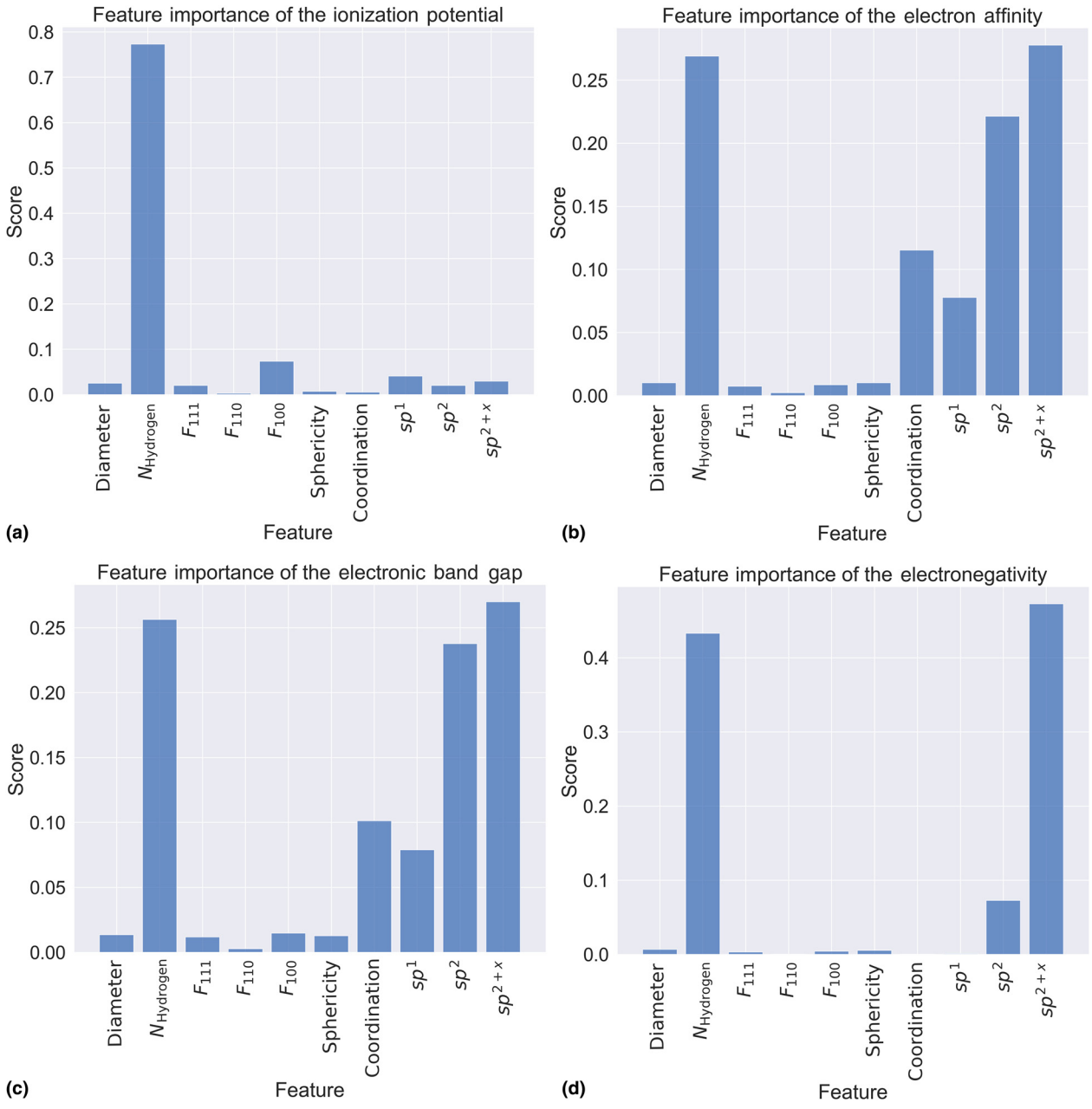


Figure 4. Histograms of the importance of the 10 features in predicting the energy of the ionization potential (a), electron affinity (b), electronic band gap (c), and electronegativity (d) of diamond nanoparticle, predicted using the RF regressor.

performance, indicating that the costly production of size-selected, monodispersed samples (whether from synthesis or post-synthesis purification) is not necessary. As the size is related to the surface-to-volume ratio (and is often used as a proxy for this more appropriate nanoscale indicator) smaller particles are still desirable, but only implicitly because they expose more surface area, not because the explicitly measured diameter is low. A distribution of sizes will not inhibit performance provided the sp^2 and sp^{2+x} atoms are maximized.

Conclusions

Using a series of machine learning methods we have demonstrated how to systematically reduce the number of dimensions in a small nanomaterials data set, to aid in visualization and interpretation, and confirm the materials features responsible for structure/property relationships. In the case of diamond nanoparticles we have confirmed the important role of surface hydrogenation, in agreement with conventional wisdom in the field, but uncovered a hidden relationship between the fraction

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press

of sp^{2+x} hybridized atoms in the particle, the electron affinity, the electronic band gap, and electronegativity. This is significant as it competes with the assumption that sp^2 hybridization (surface graphitization) drives surface charge transfer, and provide a new direction for further research into the engineering of diamond nanoparticles for a variety of applications.[22,23,24]

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1557/mrc.2019.36

References

1. A. Agrawal and A. Choudhary: Perspective: materials informatics and big data: realization of the fourth paradigm of science in materials science. *APL Mater.* **4**, 053208 (2016).
2. A. Jain, G. Hautier, S.P. Ong, and K. Persson: New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. *J. Mater. Res.* **31**, 977 (2016).
3. B. Sun, M. Fernandez, and A.S. Barnard: Statistics, damned statistics and nanoscience – using data science to meet the challenge of nanomaterial complexity. *Nano Horiz.* **1**, 89 (2016).
4. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, and Kim C: Machine learning in materials informatics: recent applications and prospects. *Comput. Mater.* **3**, 54 (2017).
5. R. Ramakrishnan and A. von Lilienfeld: Machine learning, quantum chemistry, and chemical space. *Rev. Comput. Chem.* **30**, 225 (2017).
6. B. Sun, M. Fernandez and A.S. Barnard: Machine learning for silver nanoparticle electron transfer property prediction. *J. Chem. Info. Mod.* **57**, 2413 (2017).
7. L. Ward and C. Wolverton: Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.*, **21**, 167 (2017).
8. E. Swann, B. Sun, D.M. Cleland, and A.S. Barnard: Representing molecular and materials data for unsupervised machine learning. *Molec. Simulat.* **44**, 905 (2018).
9. T. Kohonen: The self-organizing map. *Neurocomputing* **21**, 1 (1998).
10. C. Bishop: *Neural Networks for Pattern Recognition* (Oxford University Press, USA, 1995).
11. J. Gasteiger, X. X. Li, C. Rudolph, J. Sadowski, and J. Zupan: Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **116**, 46084 (1994).
12. B. Sun and A.S. Barnard: Texture based image classification for nanoparticle surface characterisation and machine learning. *J. Phys.: Mater.* **1**, 016001 (2018).
13. P. Wittek, S.C. Gao, I.S. Lim, and L. Zhao: An efficient parallel library for self-organizing maps. *J. Stat. Software*, **78**, 1 (2017).
14. A. Barnard: Nanodiamond Data Set, v1. CSIRO Data Collection (2016) doi: 10.4225/08/571F076D050B1.
15. A. Barnard: Twinned Nanodiamond Data Set, v2. CSIRO Data Collection (2018) doi: 10.25919/5be375f444e69.
16. B. Sun and A.S. Barnard: Impact of speciation on the electron charge transfer properties of nanodiamond drug carriers. *Nanoscale* **8**, 14264 (2016).
17. S. Osswald, G. Yushin, V. Mochalin, S. O. Kucheyev, and Y. Gogotsi: Control of sp^2/sp^3 carbon ratio and surface chemistry of nanodiamond powders by selective oxidation in air. *J. Am. Chem. Soc.* **128**, 11635 (2006).
18. L. Ginés, S. Mandal, A. Ahmed, C.-L. Cheng, M. Sow, and O.A. Williams: Positive zeta potential of nanodiamonds. *Nanoscale* **9**, 12549 (2017).
19. T. K. Ho: Random Decision Forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, pp. 278 (1995).
20. O.A. Williams, J. Hees, C. Dieker, W. Jäger, L. Kirste, and C.E. Nebel: Size-Dependent reactivity of diamond nanoparticles. *ACS Nano* **4**, 4824 (2010).
21. A. Nagl, S. R. Hemelaar, and R. Schirhagl: Improving surface and defect center chemistry of fluorescent nanodiamonds for imaging purposes—a review. *Anal Bioanal. Chem.* **407**, 7521 (2015).
22. K. Turcheniuk and V. Mochalin: Biomedical applications of nanodiamond (Review). *Nanotechnology* **28**, 252001 (2017).
23. A.S. Barnard: Predicting the impact of structural diversity on the performance of nanodiamond drug carriers. *Nanoscale* **10**, 8893 (2018).
24. S. Stehlik, L. Ondic, A.M. Berhane, I. Aharonovich, H.A. Girard, J.-C. Arnault, and B. Rezek: Photoluminescence of nanodiamonds influenced by charge transfer from silicon and metal substrates. *Diamond Relat. Mater.* **63**, 91 (2016).

https://doi.org/10.1557/mrc.2019.36 Published online by Cambridge University Press