ADVANCED THEORY AND SIMULATIONS
www.advtheorysimul.com

Predicting the Probability of Observation of Arbitrary Graphene Oxide Nanoflakes Using Artificial Neural Networks

*Benyamin Motevalli, Lachlan Hyde, Bronwyn L. Fox, and Amanda S. Barnard**

Although it has been well established that the stability and properties of graphene oxide nanostructure are strongly influenced by the concentration, type, and distribution of oxygen groups on the surface, there has yet to be a definitive way of predicting the thermochemical stability in advance of detailed and time-consuming experimentation or simulation. In this study, a data set of over 60 000 unique graphene oxide nanoflakes and supervised machine learning methods are used to predict the probability of observation (stability) with perfect accuracy, based on a limited set of structural features that can be controlled in advance. A decision tree is used to show how the features determine the stability, and a neural network provides an equation to predict the thermodynamic stability of virtually any configuration in minutes. This enables researchers to use machine learning as research planning tool or to assist in analyzing results from microanalysis.

1. Introduction

To date, an ever-growing body of literature has cataloged virtually all the properties of graphene oxide (GO)[1–4] for catalysis,[5] energy,[6] electrochemical,[7] sensing,[8,9] biomedicine,[10] and environmental applications.[11] Similarly, GO nanoflakes, finite sized fragments of GO, have also been studied in detail and proposed for a diverse range of applications.[12–19]

While it has been known for nearly a century that GO is a hydrophilic 2D oxidized form of graphene, with O functional groups decorating (and disrupting) the sp^2 basal plane, the structure of GO is complicated[20] but fundamentally important. The performance of GO is strongly coupled to degree and type of oxidation, and virtually all physical properties can be tuned by controlling the number and distribution of functional groups on the (upper and lower) surfaces. In addition, the type, concentration and distribution of O-groups, defects in GO, must also be considered, whether they are constructive or destructive.[7] Defects include topological features such as edges and corners, and atomistic features such as vacancies, dopants, or reconstructions. For more than 10 years, an enormous amount of scientific attention has been given to how to control the chemical structure of GO during synthesis and processing[21–26] and create a predictable and stable product based on a predetermined set of experimental conditions. Stability is important in ensuring that GO structures that are successfully produced are persistent and robust against perturbations during storage or application.[27–35] Ideally, the stability should be predictable, along with the functional properties being engineered.

Predicting the stability of GO in advance of synthesis and characterization has traditionally been the provenance of computational studies,[36–42] which were successful in comparing and ranking small sets of structures or predicting instability with respect to the environmental conditions. This approach is appropriate when the structures in the set are simple, well defined, or if the set is small. In cases where the number of unique configurations is large, the relative stability (with respect to the number, type, and distribution of O-groups and defects) must be analyzed using machine learning.[43] To date, machine learning has been used to predict various properties of graphene and GO,[44–55] but a single predictive model for the widespread prediction of the stability of arbitrary GO nanostructures has yet to be reported.

In this study, we have used interpretable supervised machine learning methods to identify the chemical and topographical features of GO nanoflakes that are most important in determining the relative stability, and report their ranking and relationship using a decision tree. We find that the stability of over 60 000

B. Motevalli
CSIRO Data61
Docklands, VIC 3008, Australia
L. Hyde
Manufacturing Futures Research Institute
Swinburne University of Technology
Hawthorn, VIC 3122, Australia
B. L. Fox
CSIRO
Clayton, VIC 3168, Australia
A. S. Barnard
School of Computing
Australian National University
Acton, ACT 2601, Australia
E-mail: amanda.s.barnard@anu.edu.au

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.202200013

DOI: 10.1002/adts.202200013

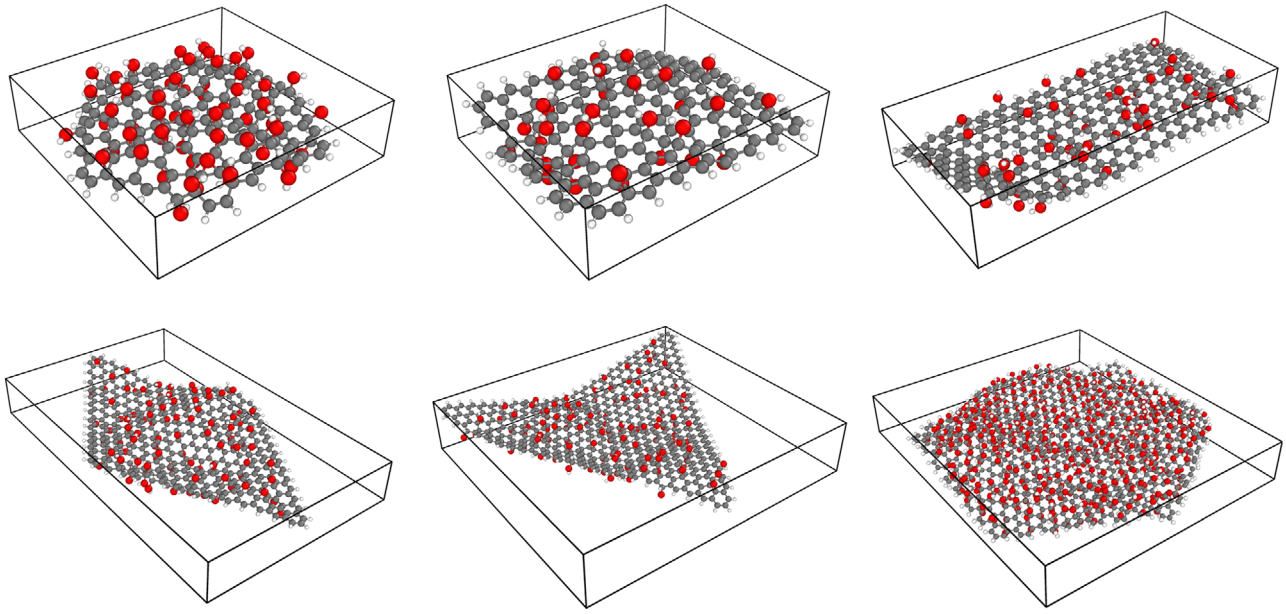

Figure 1. Examples of the graphene oxide nanoflakes contained in the data sets, contrasting different sizes, shapes, and oxygen group concentrations.

nanoflakes can be perfectly predicted by only 15 structure features, all of which are input variables that would be used to inform experiments or the generation of computational data sets. We then train an artificial neural network on these features and obtain an analytical expression capable of predicating the relative stability of GO nanoflakes prior to computationally intensive simulations or from measurements using standard microanalysis instruments.

2. Data Set and Methods

The GO data set used in this study contains 60 944 sample instances, each of which has been optimized using the density functional tight binding (DFTB) method, as described elsewhere[20,55] and in Supporting Information. This set is a combination of three data sets containing neutral, anionic, and cationic nanoflakes, each of which is available for download.[56–58] The sizes range from 320 to 2457 Å^2 with oxygen group coverage (including epoxide and ether or hydroxyl) varying from 0% to ≈52%, and morphologies including hexagonal, triangular, rectangular, and rhombic. All epoxide and ether groups were originally adsorbed as ketones before relaxation, but no ketones remained after relaxation. Variations in edges are also included by combining different ratios of armchair and zigzag edges. The data set was cleaned and outliers with a thermodynamic probability less than 10^{-8} were removed, leaving an ensemble of 60 909 unique nanoflakes. Examples of these nanoflakes are shown in **Figure 1**.

To describe each sample, 830 structural features (including the charge state) were extracted to encode a wide range of information, based on atomic, morphological, and topological descriptors.[20] After analyzing the correlation between the features, it was determined that a combination of 203 features can effectively describe the combined neutral, anionic, and cationic set, where the charges were applied adiabatically. These features are listed in Supporting Information. The target property label is thermodynamic probability of observation (P) calculated using a Boltzmann distribution and the free energy of formation at room temperature and atmospheric pressure with respect to a reservoir of water vapor. The method for calculating the environmentally dependent free energy is provided in ref. [20], beginning with the total electronic energy and using the software QuickThermo.[59] The distribution of the probability is provided in Supporting Information, showing the test/train split. This distribution is both imbalanced and bimodal (due to a larger number of smaller nanoflakes in the set and a distinct subset of much larger nanoflakes), indicating that standardization and stratification are required, as described in the Results. This is an important step to avoid information leakage, where the range of the testing set impacts the range of the training set. Stratification ensures the distribution is the same at the outset, so it is not necessary to standardize the testing and training sets independently.[60,61]

Regression is a type of supervised learning where the target labels are also provided with the features. In this study, we have used the non-linear, non-parametric decision tree regressor (DTR) to train model that predicts the relationship between the structural features and the thermodynamic probability of observation based on simple decision rules inferred from the features. Decision trees are trained by recursively splitting the data and are able to handle multi-output problems. They are simple to understand, and an explanation for the condition is easily explained by boolean logic. Advantages of a decision tree regressor is that it requires little data preparation, it can be validated using statistical tests, and it can provide feature importance rankings and can be visualized as a decision tree. The hyper-parameters of the DTR were optimized using a grid search and applied using ten-fold stratified cross-validation and a 20/80 test/train split. Disadvantages include possible instability with respect to small variation in the data, locally optimal decisions at nodes dominating since they are based on heuristic algorithms (given an optimal decision

tree is known to be NP-complete[62]), biased trees being created if some classes dominate, and a propensity toward over-fitting (they do not generalize well). This makes hyper-parameter optimization very important. For these reasons, the DTR was chosen for its intuitive interpretability, but we also calculated the learning curve to test for convergence of the training and cross validation scores to evaluate the accuracy and generalizability.

As we will see, once the important features have been confirmed, we used a multi-layer perceptron (MLP, a type of artificial neural network) to obtain an analytical expression for rapid prediction of the probability of arbitrary GO nanoflakes using a limited set of variables. An advantage of using a neural network is that the model can be extracted analytically, in the form:

$$o_i = f\left(\Sigma_i a_i x_i + b\right) \tag{1}$$

where x_i is the output o of neuron i in the previous layer, a_i is the weight of neuron i in the previous layer which is determined during training, b is the bias determined during training, and f is the activation function. The output of neuron n in layer l of the neural network is given by:

$$o_n^l = \left[f\left(\Sigma_m a_{nm}^l\left[\ldots\left[f\left(\Sigma_j a_{kj}^2\left[f\left(\Sigma_i a_{ji}^1 x_i + b_j^1\right)\right] + b_k^2\right)\right]\ldots\right] + b_n^l\right)\right] \tag{2}$$

where x_i is the i^{th} input feature, a_{hk}^l is the weight in neuron h of layer l of the output of neuron k from layer $l-1$ in layer l, b_p^q is the bias added to neuron p in layer q, and once again f is the activation function. The time complexity of the MLP with back propagation as given by the scikit-learn package[63] is $\mathcal{O}(n \times m \times h^l \times i)$, where n is the number of training samples, m is the number of features, h^l is the number of neurons in each layer, l is the number of layers, and i is the number of iterations. Training the neural network involves calculating the error and back propagating this error to each of the neurons, which updates the coefficients and biases. The MLP regressor uses the log-loss function and can be evaluated using any metric. This also included ten-fold stratified cross-validation with a 20/80 test/train split, and evaluation using training, testing, and cross-validation scores, and the learning curve.

3. Discussion of Results

To begin with, we trained a DTR to identify the most important structural features responsible for predicting the thermodynamic probability of observation, P. All structural features were normalized using a standard scaler which removes the mean and scales to unit variance, and the hyper-parameters of the model were optimized using a random grid search of 1000 iterations. The optimal DTR was found to require: criterion="mae", max_leaf_nodes=None, min_impurity_decrease=10^{-7}, max_depth=10, splitter="best" , max_features="auto", min_samples_leaf=5, and min_samples_split=5, with a random state of 42. Training this model on the 60 909 instances and in the 203 dimensional feature space gave a perfect result of an R^2_{train}=1.0, R^2_{train}=1.0, and cross-validation score 1.0 ± 0.00003. It is rare to achieve such a good result, which is partly due to appropriate model selection, partly due to the identification of ideal hyper-parameters, and partly due to the existing of a strong underlying relationship between the target label and the data. This is more common in computational data sets than experimental data sets, but is still difficult to achieve. The training and testing plots are shown in **Figures 2**a and 2b, respectively. The decision tree is provided in Supporting Information, but has only been used here for the purposes of feature selection.

The learning curve shown in **Figure 3**a, which shows the model performance with increasing amounts of data, testifies to the lack of under-fitting or over-fitting, but to determine how many of the features were contributing to the model, we performed recursive feature elimination with ten-fold cross validation, and as shown in Figure 3b, there is no change beyond the optimal number of 15 features. In Figure 3c, we can also see the feature importance profile, which provides a ranking of how influential each of these structural features are in determining P. Here, we can see that the profile rapidly decays and that the importance of the 15th feature of the 203 (the oxygen concentration) is effectively zero. In principle, any interpretable feature selection method could be used, including methods such as least absolute shrinkage and selection operator regularization (LASSO)[64]; however, some that use shrinkage (like LASSO) have the disadvantage of basing the dimension reduction on the feature variance, rather than the relative importance to the label prediction. Low variance features will be set to zero by the L1 regularization (which adds a penalty equal to the absolute value of the magnitude of coefficients), regardless of whether there is a strong correlation to the target. Methods based as RFE using feature importance profiles, or explainable AI,[65] ensure that features with a strong relationship to the target are retained, even if the numerical variance is low.

The 15 most important features are listed in **Table 1**. These structural features are intuitive (which they may not have been if a shrinkage method such as LASSO had been used), but do not include the fraction of aromatic (sp^2-bonded) carbon, or defects related to ether groups,[55] which are also often considered to be energetically important. On a case-by-case basis broken bond defects may impact the formation energy significantly, but these results indicated that in a collection of over 60 000 a defective nanoflake is no more or less probable than one with the C–C bonds intact, as the only feature related to defect formation (ring_density_7_reducible_all, drr_7) is of low rank. In general, the probability of observation is determined by the size, edge passivation, and the type and distribution of oxygen-containing groups, but not their concentration. This is counter intuitive, since we know that the total energy of the nanoflakes will be affected by the oxygen concentration, however, according to the data, the quantity oxygen is not as important as the it is clustered or separated a over the nanoflake . This is good news for manufacturers as it suggests that small variations in the oxygen concentration can be tolerated, provided the coverage is controlled.

These results also provide a remarkable opportunity. All of the most important structural features identified by the DTR relate to input design parameters, and do not require electronic structure simulations (or any other simulations) to calculate them. The fraction of sp^2-bonded C atoms and many other structural features in the set, all require the nanoflake structures to be fully relaxed, but the features listed in Table 1 can be determined for

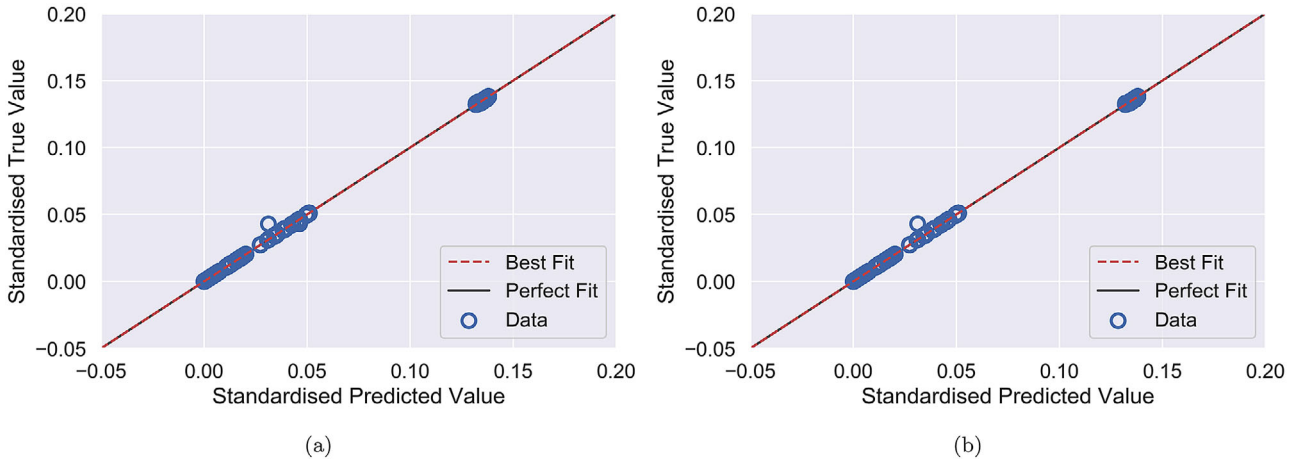

Figure 2. a) The training results for the decision tree regressor trained on 48727 GO nanoflakes using 203 structural features, and b) the testing results (12182 nanoflakes and 203 features).

Table 1. Important features from the optimized decision tree regressor in a 203 dimensional feature space, ranked in order of influence in determining the thermodynamic probability of observation, *P*. Note that the H–C_-coordination and C–H_-coordination are not the inverse of the same feature, as H–C_-coordination omits H–O bonds and ignores C–O bonds, whereas C–H_-coordination omits C–O bonds and ignores H–O bonds. No hydrogen atoms were adsorbed in-plane during data generation, so H–C_-coordination omitting H–O bonds can only reside on the edges.

Rank	feature	Interpretation	Symbol
1	all_agent_group_count	The total number of hydroxyls, ethers, and carboxyls over a given area	N_{groups}
2	H–C_-coordination	The fraction of H atoms bonded to C atoms, reflecting the degree of edge passivation	HC_{coord}
3	min_diameter	The minimum diameter, which is reflected in the anisotropy	D_{min}
4	area	The total area of the GO nanoflake, reflecting the overall size	A
5	norm_Iyy_hydro_area	The moment of inertia of the normalized density distribution of hydroxyl groups around the *y*-axis normalized by the area, reflecting the distribution of OH	I_{yy}(OH)
6	H_concentration	The concentration of H atoms	c(H)
7	C–H_-coordination	The fraction of C atoms bonded to H atoms, reflecting the degree of edge passivation	CH_{coord}
8	hydroxyl_concentration	The fraction of oxygen groups that are hydroxyl (as opposed to epoxide, ether, or carboxyl)	c(OH)
9	ZZ_edge	The fraction of zigzag edges, which reflects the fraction of armchair edges, as they sum to 1	ZZ
10	C_concentration	The concentration of C atoms	c(H)
11	norm_Ixx_O_area	The moment of inertia of the normalized density distribution of oxygen atoms around *x*-axis normalized by area, reflecting the distribution of O	I_{xx}(O)
12	ring_density_7_reducible_all	Density of reducible (all) 7 membered rings formed when an epoxide breaks a C–C bond to form an ether, reflecting the density of ethers	drr_7
13	density_loc_hydro_mean	The average local density of hydroxyl groups, reflecting the degree of clustering of OH	ρ(OH)
14	max_diameter	The maximum diameter, which is reflected in the anisotropy	D_{max}
15	O_concentration	The concentration of O atoms	c(O)

un-relaxed nanoflake structures, or based on hypothetical combinations. To confirm that these 15 structural features are sufficient to predict *P*, we re-optimized and trained a DTR on the entire 60 909 GO set using only these structural features. Once again, the model was optimized using a random grid search, but returned the same set of hyper-parameters. The reduced set gave an $R^2_{train} = R^2_{test}$=1.0, and cross-validation score 0.999 ± 0.003. This confirms that this limited set of input design parameters can be used to predict the probability of a given GO nanoflake prior to any simulation. The hyper-parameters for this model, in addition to plots of the results, is contained in Supporting Information.

The actual tree explaining how the model came to these results is provided in Supporting Information (due to the size), where we can see it has a depth of 16 and initially splits based on the area, and then again on the area and H-concentration, before splitting on the most important feature at a depth of 4. It is not essential that decision trees split in order of feature importance, as is the case here. Readers will also note that there are "outliers" in the DTR prediction, though the disagreement is less than 1%. These instances are hexagonal nanoflakes 4.8 nm in diameter, with armchair edges and 16% oxygen concentration. They have significant out-of-plane distortions, and a higher than normal fraction of "strained" bonds, where angles are distorted (more than 140 degrees).

Decision trees are generally not specific enough to make a specific prediction. To increase the utility of these results, we

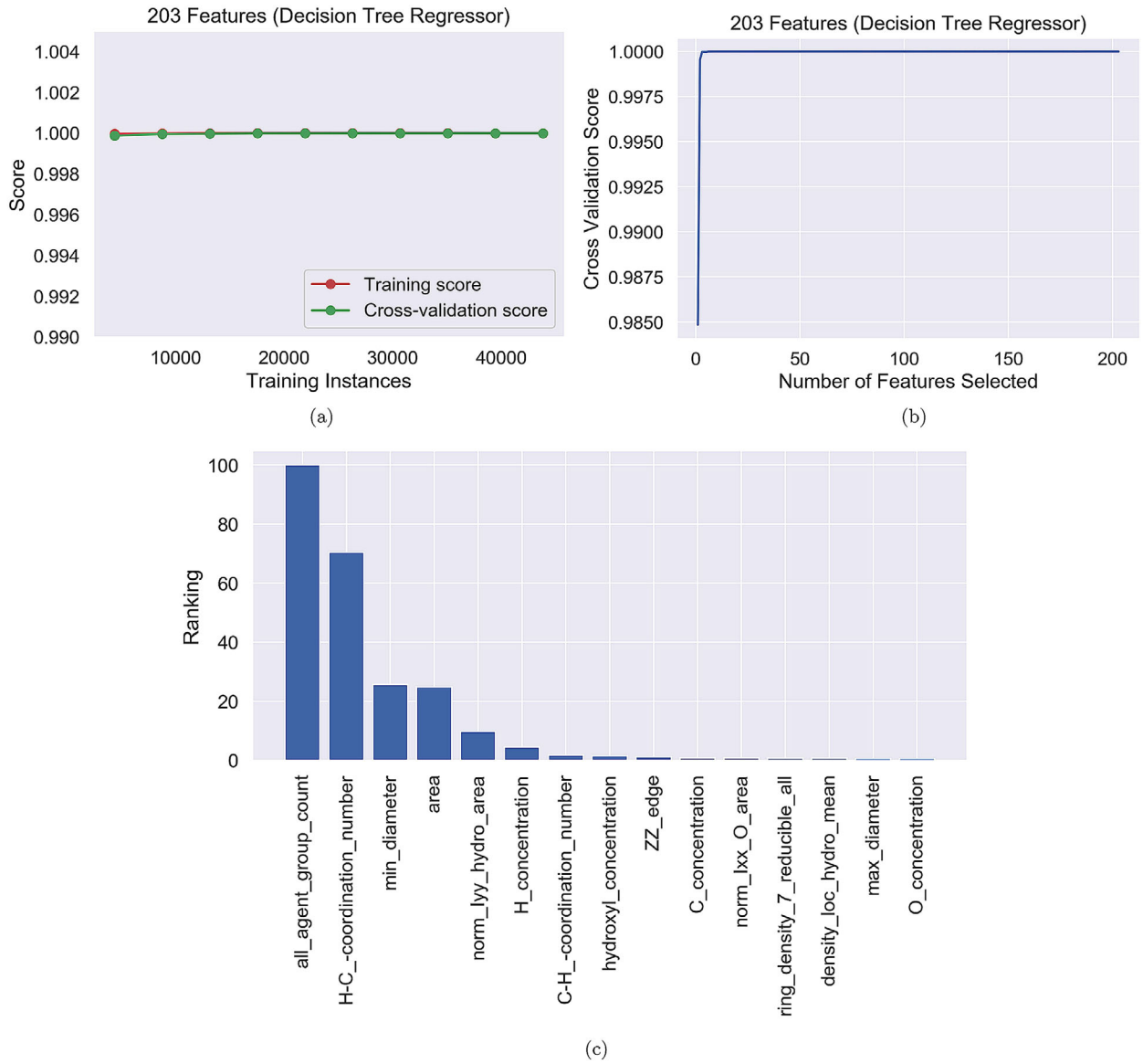

Figure 3. a) The learning curve showing a perfect result with no over-fitting or under-fitting, b) the feature elimination plot showing no change in the cross-validation score beyond an optimal number of features of 15, and c) the feature importance ranking of the top 15 structural features in order of influence.

trained an artificial neural network (ANN) using the reduced feature set to predict P. The ANN was optimized using a random grid search over 1000 hyper-parameter combinations allowing up to three hidden layer and up to 15 neurons per hidden layer, resulting in the hyper-parameters: solver="adam", activation = "tanh", max_iter=257, alpha = 0.008, learning_rate="adaptive", hidden_layer_sizes=(8,), early_stopping=True, with a random state of 42. The tolerance for the optimization was 10^{-4}, the numerical stability was 10^{-8}, and samples were shuffled in each iteration. This means we have one hidden layer with eight neurons. In this case, since we seek an analytical expression that can be used to predict P directly, we have used the un-scaled feature values, and obtained an $R^2_{train} = R^2_{test}$=0.995, and cross-validation score 0.994 ± 0.005. The resultant weights and biases give the following expression for predicting the probability of observation of any small GO nanoflake with a total oxygen concentration under 50%:

$$P = -0.07158h_1 - 0.1319h_2 + 0.0010h_3 - 0.0222h_4 - 0.1053h_5 - 0.0083h_6 + 0.0716h_7 - 0.0807h_8 + 0.1192 \quad (3)$$

with

$$h_1 = f(0.00048N_{groups} + -0.00472HC_{coord} - 0.01325D_{min} - 0.01397A - 0.04451I_{yy}(\mathrm{OH}) - 0.00429c(\mathrm{H}) - 0.02898CH_{coord} - 0.00229c(\mathrm{OH}) + 0.00196ZZ + 0.00163c(\mathrm{C}) + 0.00757I_{xx}(\mathrm{O}) - 0.00353drr_7$$

$$-0.04675\rho(\mathrm{OH}) - 0.02241 D_{max} + 0.00441 c(\mathrm{O}) + 0.31229) \quad (4)$$

$$h_2 = f(-0.00930 N_{groups} + 0.01569 HC_{coord} - 0.08178 D_{min} - 0.05628 A - 0.12989 I_{\gamma\gamma}(\mathrm{OH}) + 0.01784 c(\mathrm{H}) - 0.08519 CH_{coord} - 0.02502 c(\mathrm{OH}) + 0.03753 ZZ + 0.02478 c(\mathrm{C}) + 0.01639 I_{xx}(\mathrm{O}) - 0.01584 drr_7 + 0.00304\rho(\mathrm{OH}) + 0.00023 D_{max} + 0.00181 c(\mathrm{O}) + 0.70646) \quad (5)$$

$$h_3 = f(0.00021 N_{groups} - 0.00007 HC_{coord} + 0.00024 D_{min} + 0.00014 A + 0.00003 I_{\gamma\gamma}(\mathrm{OH}) - 0.00001 c(\mathrm{H}) - 0.00021 CH_{coord} - 0.00012 c(\mathrm{OH}) - 0.00009 ZZ + 0.00014 c(\mathrm{C}) + -0.00008 I_{xx}(\mathrm{O}) - 0.00019 drr_7 - 0.00191\rho(\mathrm{OH}) - 0.00100 D_{max} - 0.00022 c(\mathrm{O}) + 0.04104) \quad (6)$$

$$h_4 = f(0.00279 N_{groups} - 0.00269 HC_{coord} - 0.00036 D_{min} - 0.00368 A - 0.00973 I_{\gamma\gamma}(\mathrm{OH}) - 0.00170 c(\mathrm{H}) - 0.00838 CH_{coord} - 0.00319 c(\mathrm{OH}) + 0.00229 ZZ + 0.00331 c(\mathrm{C}) + 0.00092 I_{xx}(\mathrm{O}) - 0.00417 drr_7 + 0.00597\rho(\mathrm{OH}) + 0.01535 D_{max} - 0.00207 c(\mathrm{O}) - 0.29152) \quad (7)$$

$$h_5 = f(0.00326 N_{groups} - 0.00894 HC_{coord} + 0.07964 D_{min} + 0.09312 A + 0.08193 I_{\gamma\gamma}(\mathrm{OH}) - 0.00487 c(\mathrm{H}) + 0.08820 CH_{coord} + 0.03253 c(\mathrm{OH}) + -0.02915 ZZ - 0.01515 c(\mathrm{C}) - 0.02138 I_{xx}(\mathrm{O}) + 0.02707 drr_7 + 0.04232\rho(\mathrm{OH}) + 0.05010 D_{max} + 0.00135 c(\mathrm{O}) + 0.07012) \quad (8)$$

$$h_6 = f(0.00170 N_{groups} - 0.00149 HC_{coord} + 0.00018 D_{min} - 0.00131 A - 0.00452 I_{\gamma\gamma}(\mathrm{OH}) - 0.00079 c(\mathrm{H}) - 0.00433 CH_{coord} - 0.00165 c(\mathrm{OH}) + 0.00090 ZZ + 0.00176 c(\mathrm{C}) + 0.00026 I_{xx}(\mathrm{O}) - 0.00242 drr_7 - 0.00120\rho(\mathrm{OH}) + 0.00122 D_{max} - 0.00132 c(\mathrm{O}) - 0.17209) \quad (9)$$

$$h_7 = f(-0.00911 N_{groups} + 0.00746 HC_{coord} - 0.00362 D_{min} + 0.00636 A + 0.02479 I_{\gamma\gamma}(\mathrm{OH}) + 0.00399 c(\mathrm{H}) + 0.01913 CH_{coord} + 0.00438 c(\mathrm{OH}) - 0.00184 ZZ - 0.00796 c(\mathrm{C}) - 0.00179 I_{xx}(\mathrm{O}) + 0.01081 drr_7 - 0.02900\rho(\mathrm{OH}) - 0.00796 D_{max} + 0.00556 c(\mathrm{O}) + 0.34999) \quad (10)$$

$$h_8 = f(-0.00214 N_{groups} + 0.00154 HC_{coord} - 0.02285 D_{min} - 0.02306 A - 0.05680 I_{\gamma\gamma}(\mathrm{OH}) + 0.00219 c(\mathrm{H}) - 0.03819 CH_{coord} - 0.01243 c(\mathrm{OH}) + 0.01419 ZZ + 0.01114 c(\mathrm{C}) + 0.01245 I_{xx}(\mathrm{O}) - 0.00875 drr_7 - 0.04288\rho(\mathrm{OH}) - 0.04409 D_{max} + 0.00197 c(\mathrm{O}) + 0.52445) \quad (11)$$

where f is the hyperbolic tangent (tanh) activation function,

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (12)$$

The model fits and learning curves for the ANN are shown in **Figure 4**, where we can see that the ANN can handle the outliers effectively. Reducing the number of neurons in the hidden layer to seven resulted in a significant reduction in performance, though with $R^2_{train} = R^2_{test} = 0.995$, and cross-validation score 0.993 ± 0.003. While these scores are not significantly different and a reduction of one neuron would simplify the system of equations, a comparison of the predicted value and the ground truth shows it is a fortuitous cancellation of errors and that the residuals have increased significantly (see Supporting Information). A linear regressor using the same 15 important features returned $R^2_{train} = R^2_{test} = 0.948$ cross-validation score 0.868 ± 0.482, introducing only minor under-fitting, but severely undermining the generalizability by introducing significant over-fitting.

Although this is a complicated expression, particularly given that f is the tanh activation function, it is possible to code this into any program language or sophisticated spreadsheet package. With these expressions, it is possible to predict whether a given GO nanoflake structure is likely to have a high or low probability of observation (be stable or unstable), without needing to simulate it or a range of alternatives to compare to. This provides the opportunity for rapid pre-screening of candidates prior to high-throughput computational discovery and design, reducing the risk of identifying high-performing configurations that will be difficult to produce in reality.[66] It also facilitates more targeted simulation strategies that reduce the number of simulations required to find low-energy configurations that converge rapidly to their ground state, making computational studies less expensive and energy intensive. This is particularly important given recent concerns about the effects of this power consumption on the environment, as high-performance computing platforms typically use as much energy

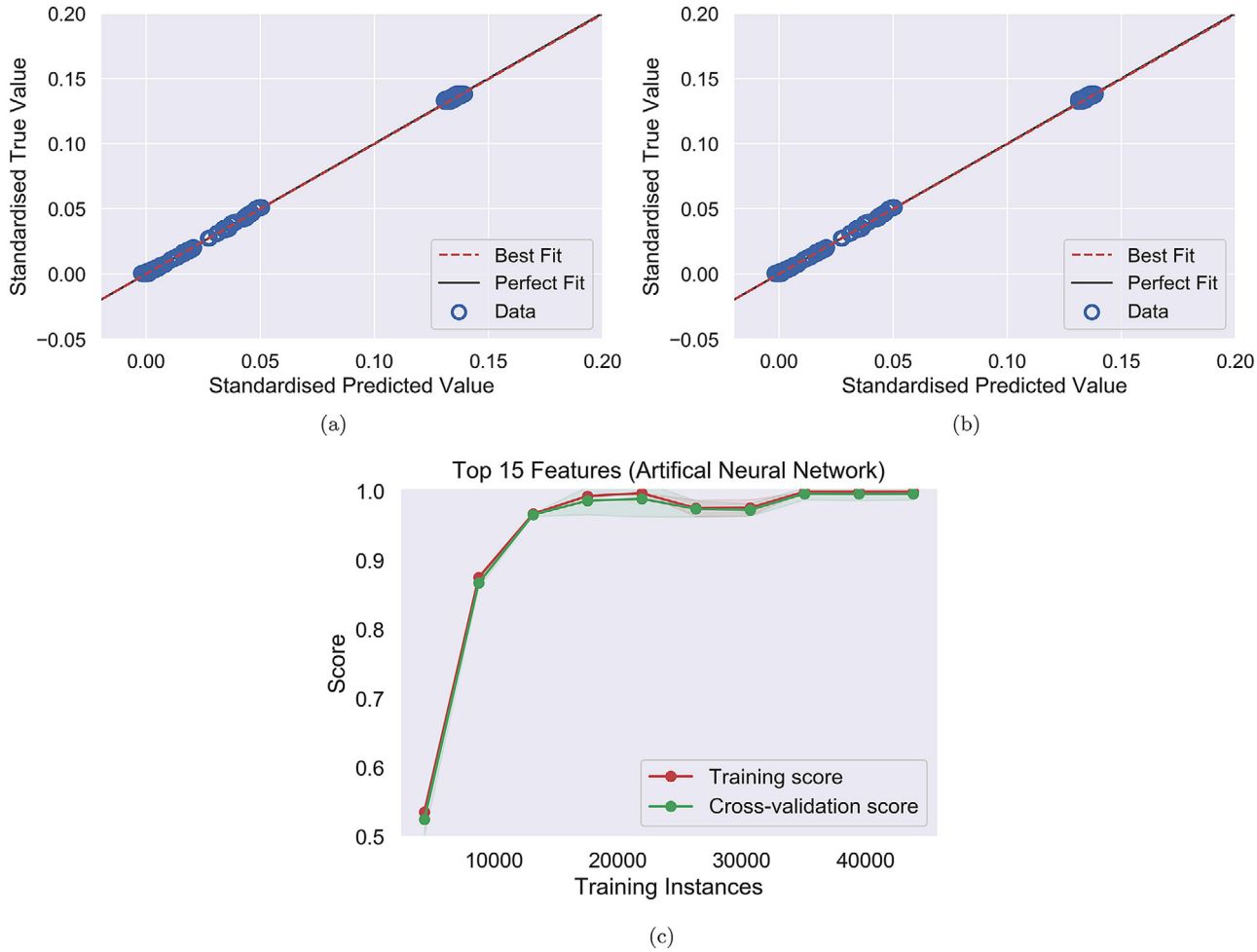

Figure 4. a) The training results for the artificial neural network trained on 48727 GO nanoflakes using the top 15 structural features, and b) the testing results (12182 nanoflakes and 15 features), and c) the learning curve showing a perfect result with no over-fitting and minimal under-fitting, provided sufficient data are over 35 000 instances.

as a small city.[67] New processors such as tensor processing units designed specifically for machine learning are even less energy efficient, so hardware optimization is likely to drive even higher energy needs.[68] Even the collective energy consumption of the huge number of personal computers currently training and running ML models is staggering[69] and has led to calls for "Green AI".[70]

These results demonstrate that decisions made regarding the experimental design in laboratory studies have a salient influence on the stability of GO nanoflakes. Only stable structures have a high probability of observation and may be expected to survive the rigors of processing or purification strategies, preserving the structure and composition intended during synthesis.[21–26] By tuning the variables identified here, it may also be possible to drive synthesis into a more stable region of the configuration space without the need for excessive trial and error.[27–35] Stability is one of the only areas with emerging standards for nanomaterials, including graphene and graphene oxide.[71] Stable GO structures with a higher probability of observation will also be less prone to degradation, with limited aging effects,[72] and has so far limited the supply chain length for graphene products. This could be an advantage to manufacturers who aim to provide reliable and reproducible products to their customers.

4. Conclusions

To summarize the results, we have shown using a data set of over 60 000 unique graphene oxide nanoflakes created using electronic structure simulations, and a series of machine learning methods, that the probability of observation (stability) can be predicted in advance of simulations or experimentation based on a limited set of human interpretable structural features. A decision tree has been generated to show how these features determine the stability, and a neural network has been used to provide an analytical expression that can be used without needing to perform thousands of simulations, or in concert with measurements taken in the lab. The models all have perfect accuracy and generalizability and can be used with confidence to predict the thermodynamic stability of virtually any configuration in minutes. This enables researchers to use machine learning as research planning tool, as well as a powerful way to analyze new results.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Computational resources for this project were supplied by the National Computing Infrastructure (NCI) national facility under partner Grant p00.

Open access publishing facilitated by Australian National University, as part of the Wiley - Australian National University agreement via the Council of Australian University Librarians.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in CSIRO Data Collection at https://doi.org/10.25919/5e30b44a7c948, https://doi.org/10.25919/5e30a9cf118cf and https://doi.org/10.25919/5e30a9cf90439, reference number [1].

Keywords

design, graphene oxide, machine learning, manufacture

Received: January 9, 2022
Revised: February 1, 2022
Published online: February 20, 2022

[1] P. P. Brisebois, M. Siaj, *J. Mater. Chem. C* **2020**, *8*, 1517.
[2] S. Stankovich, D. A. Dikin, R. D. Piner, K. A. Kohlhaas, A. Kleinhammes, Y. Jia, Y. Wu, S. T. Nguyen, R. S. Ruoff, *Carbon* **2007**, *45*, 1558.
[3] Y. Zhu, S. Murali, W. Cai, X. Li, J. W. Suk, J. R. Potts, R. S. Ruoff, *Adv. Mater.* **2010**, *22*, 3906.
[4] F. Perrozzi, S. Prezioso, L. Ottaviano, *J. Phys.: Condens. Matter* **2014**, *27*, 013002.
[5] H. Ahmad, M. Fan, D. Hui, *Compos. Part B: Eng.* **2018**, *145*, 270.
[6] D. Li, M. B. Müller, S. Gilje, R. B. Kaner, G. G. Wallace, *Nat. Nanotechnol.* **2008**, *3*, 101.
[7] D. Chen, H. Feng, J. Li, *Chem. Rev.* **2012**, *112*, 6027.
[8] N. Sharma, V. Sharma, Y. Jain, M. Kumari, R. Gupta, S. K. Sharma, K. Sachdev, *Macromol. Symp.* **2017**, *376*, 1700006.
[9] S. J. Rowley-Neale, E. P. Randviir, A. S. Abo Dena, C. E. Banks, *Appl. Mater. Today* **2018**, *10*, 218.
[10] T. P. D. Shareena, D. McShan, A. K. Dasmahapatra, P. B. Tchounwou, *Nano-Micro Letters* **2018**, *10*, 53.
[11] F. Li, X. Jiang, J. Zhao, S. Zhang, *Nano Energy* **2015**, *16*, 488.
[12] A. M. Silva, M. S. Pires, V. N. Freire, E. L. Albuquerque, D. L. Azevedo, E. W. S. Caetano, *J. Phys. Chem. C* **2010**, *114*, 17472.
[13] M. Nair, D. Nancy, A. G. Krishnan, G. S. Anjusree, S. Vadukumpully, S. V. Nair, *Nanotechnology* **2015**, *26*, 161001.
[14] Y. Zhao, W. Ran, J. He, Y. Huang, Z. Liu, W. Liu, Y. Tang, L. Zhang, D. Gao, F. Gao, *Small* **2015**, *11*, 1310.
[15] G. S. Nayak, R. Zybala, R. Kozinski, M. Woluntarski, R. Telle, K. Schickle, *Mater. Lett.* **2018**, *225*, 109.
[16] V. Castagnola, W. Zhao, L. Boselli, M. C. L. Giudice, F. Meder, E. Polo, K. R. Paton, C. Backes, J. N. Coleman, K. A. Dawson, *Nat. Commun.* **2018**, *9*, 1577.
[17] X. Ge, J. Li, R. Luo, C. Zhang, J. Luo, *ACS Appl. Mater. Interfaces* **2018**, *10*, 40863.
[18] R. Di Santo, L. Digiacomo, S. Palchetti, V. Palmieri, G. Perini, D. Pozzi, M. Papi, G. Caracciolo, *Nanoscale* **2019**, *11*, 2733.
[19] S. Taniselass, M. M. Arshad, S. C. Gopinath, M. M. Ramli, *J. Colloid Interface Sci.* **2020**, *577*, 345.
[20] B. Motevalli, A. J. Parker, B. Sun, A. S. Barnard, *Nano Futures* **2019**, *3*, 045001.
[21] Y. Zhu, S. Murali, W. Cai, X. Li, J. W. Suk, J. R. Potts, R. S. Ruoff, *Adv. Mater.* **2010**, *22*, 3906.
[22] D. R. Dreyer, S. Park, C. W. Bielawski, R. S. Ruoff, *Chem. Soc. Rev.* **2010**, *39*, 228.
[23] D. C. Marcano, D. V. Kosynkin, J. M. Berlin, A. Sinitskii, Z. Sun, A. Slesarev, L. B. Alemany, W. Lu, J. M. Tour, *ACS Nano* **2010**, *4*, 4806.
[24] A. M. Dimiev, J. M. Tour, *ACS Nano* **2014**, *8*, 3060.
[25] R. K. Singh, R. Kumar, D. P. Singh, *RSC Adv.* **2016**, *6*, 64993.
[26] P. V. Ranjan, S. Agrawal, A. Sinha, T. R. Rao, J. Balakrishnan, A. D. Thakur, *Sci. Rep.* **2018**, *8*, 12007.
[27] S. Zhou, A. Bongiorno, *Sci. Rep.* **2013**, *3*, 2484.
[28] F. Tu, S. Liu, G. Jin, G. Yan, C. Pan, *Powder Technol.* **2013**, *249*, 146.
[29] I. Chowdhury, M. C. Duch, N. D. Mansukhani, M. C. Hersam, D. Bouchard, *Environ. Sci. Technol.* **2013**, *47*, 6288.
[30] S. Eigler, S. Grimm, A. Hirsch, *Chem. Eur. J.* **2014**, *20*, 984.
[31] C.-N. Yeh, K. Raidongia, J. Shao, Q.-H. Yang, J. Huang, *Nat. Chem.* **2015**, *7*, 166.
[32] M. M. Gudarzi, *Langmuir* **2016**, *32*, 5058.
[33] N. Ye, Z. Wang, S. Wang, H. Fang, D. Wang, *Environ. Sci. Pollut. Res.* **2018**, *25*, 10956.
[34] J. Nuncira, L. M. Seara, R. D. Sinisterra, V. Caliman, G. G. Silva, *Fullerenes, Nanotubes Carbon Nanostruct.* **2020**, *28*, 407.
[35] R. Patil, D. Marathe, S. P. Roy, D. Ray, V. K. Aswal, P. K. Jha, P. Bahadur, S. Tiwari, *Mater. Sci. Eng: C* **2020**, *109*, 110559.
[36] R. J. W. E. Lahaye, H. K. Jeong, C. Y. Park, Y. H. Lee, *Phys. Rev. B* **2009**, *79*, 125435.
[37] A. S. Barnard, I. K. Snook, *J. Mater. Chem.* **2010**, *20*, 10459.
[38] L. Wang, Y. Y. Sun, K. Lee, D. West, Z. F. Chen, J. J. Zhao, S. B. Zhang, *Phys. Rev. B* **2010**, *82*, 161406.
[39] H. Shi, A. S. Barnard, I. K. Snook, *J. Mater. Chem.* **2012**, *22*, 18119.
[40] H. Shi, A. S. Barnard, I. K. Snook, *Phys. Chem. Chem. Phys.* **2013**, *15*, 4897.
[41] H. Shi, L. Lai, I. K. Snook, A. S. Barnard, *J. Phys. Chem. C* **2013**, *117*, 15375.
[42] F. Mouhat, F.-X. Coudert, M.-L. Bocquet, *Nat. Commun.* **2020**, *11*, 1566.
[43] A. S. Barnard, B. Motevalli, A. J. Parker, J. M. Fischer, C. A. Feigl, G. Opletal, *Nanoscale* **2019**, *11*, 19190.
[44] M. Fernandez, A. S. Barnard, *ACS nano* **2015**, *9*, 11980.
[45] M. Fernandez, H. Shi, A. S. Barnard, *J. Chem. Inf. Model.* **2015**, *55*, 2500.
[46] M. Fernandez, H. Shi, A. S. Barnard, *Carbon* **2016**, *103*, 142.
[47] M. Fernandez, J. I. Abreu, H. Shi, A. S. Barnard, *ACS Comb. Sci.* **2016**, *18*, 661.
[48] X. Yan, A. Sedykh, W. Wang, X. Zhao, B. Yan, H. Zhu, *Nanoscale* **2019**, *11*, 8352.
[49] Z. Zhang, Y. Hong, B. Hou, Z. Zhang, M. Negahban, J. Zhang, *Carbon* **2019**, *148*, 115.
[50] Y. Dong, C. Wu, C. Zhang, Y. Liu, J. Cheng, J. Lin, *npj Comput. Mater.* **2019**, *5*, 1.
[51] M. Yamawaki, M. Ohnishi, S. Ju, J. Shiomi, *Sci. Adv.* **2018**, *4*, 6 eaar4192.
[52] S. Masubuchi, T. Machida, *npj 2D Mater. Appl.* **2019**, *3*, 4.
[53] J. Wan, J.-W. Jiang, H. S. Park, *Carbon* **2020**, *157*, 262.

[54] T. Zhang, Q. Liu, Y. Dan, S. Yu, X. Han, J. Dai, K. Xu, *Opt. Express* **2020**, *28*, 18899.
[55] B. Motevalli, B. Sun, A. S. Barnard, *J. Phys. Chem. C* **2020**, *124*, 7404.
[56] A. Barnard, B. M. Soumehsaraei, B. Sun, L. Lai, *CSIRO Data Collect.* **2019**.
[57] A. Barnard, B. M. Soumehsaraei, B. Sun, L. Lai, *CSIRO Data Collect.* **2019**.
[58] A. Barnard, B. M. Soumehsaraei, B. Sun, L. Lai, *CSIRO Data Collect.* **2019**.
[59] B. M. Soumehsaraei, A. Barnard, *CSIRO Data Collect.* **2019**.
[60] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, *Chem. Mater.* **2020**, *32*, 4954.
[61] A. S. Barnard, *Matter* **2020**, *3*, 22.
[62] L. Hyafil, R. L. Rivest, *Inf. Process. Lett.* **1976**, *5*, 15.
[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[64] F. Santosa, W. W. Symes, *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 307.
[65] T. Liu, A. S. Barnard, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 035034.
[66] A. Zunger, *Nature* **2019**, *566*, 447.
[67] M. Halper, *ACM News* **2015**.
[68] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, et al., *SIGARCH Comput. Archit. News* **2017**, *45*, 1.
[69] E. García-Martín, C. F. Rodrigues, G. Riley, H. Grahn, *J. Parallel Distrib. Comput.* **2019**, *134*, 75.
[70] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, *Commun. ACM* **2020**, *63*, 54.
[71] I. N. T. Committee, Iso/dts 21357, nanotechnologies — evaluation of the mean size of nano-objects in liquid dispersions by static multiple light scattering (smls), Technical Report first edition, International Organization for Standardization (ISO), https://www.iso.org/standard/70759.html.
[72] V. B. Mohan, R. Brown, K. Jayaraman, D. Bhattacharyya, *Mater. Sci. Eng: B* **2015**, *193*, 49.