The representative structure of graphene oxide nanoflakes from machine learning

Benyamin Motevalli, Amanda J Parker, Baichuan Sun and Amanda S Barnard
Data61 CSIRO, Docklands, Australia
E-mail: benyamin.motevalli@data61.csiro.au
Keywords: machine learning, graphene oxide, nanoparticles, archetypal analysis, clustering
Supplementary material for this article is available online

Abstract
In this paper we revisit the structure of graphene oxide, and determine the pure and truly representative structures for graphene nanoflakes using machine learning. Using 20 396 random configurations relaxed at the electronic structure level, we observe the presence of hydroxyl, ether, double bonds, aliphatic (cyclohexane) disruption, defects and significant out-of-plane distortions that go beyond the Lerf–Klinowski model. Based on a diverse list of 224 chemical, structural and topological features we identify 25 archetypal ‘pure’ graphene oxide structures which capture all of the complexity and diversity of the entire data set; and three prototypes that are the truly representative averages in 224-dimensional space. Together these 28 structures, which are shown to be largely robust against changes in thermochemical conditions modeled using ab initio thermodynamics, can be downloaded and used collectively as a small data set for with a fraction of the computational cost in future work, or independently as an exemplar of graphene oxide with the required oxidation.

1. Introduction
What is the actual structure of graphene oxide nanoflakes? There have been countless studies on the properties of graphene oxide (GO) [1–4] for electrochemical [5], bio-medicine [6], sensing [7], catalysis [8], energy [9] and environmental applications [10]. It has been well established that the performance of GO is strongly coupled to degree and type of oxidation, and virtually all of physical properties can be tuned by controlling the number and distribution of functional groups on the (upper and lower) surfaces, with far fewer studies focussing on what the representative structure look like, and none delivering atomic coordinates that can be used by others for ongoing work.

It has been known for nearly a century that GO is a hydrophillic two dimensional oxidized form of graphene, with O functional groups decorating (and disrupting) the sp² basal plane, ranging in size from a few nm to mm. The first model of the structure of GO was proposed by Hofmann and Holst, [11] such that oxygen was bound to a hexagonal carbon sheet by epoxy (1, 2-ether) with a formula of C₂O. This model was revised by a series of authors over the last 80 years, taking into account the wrinkling of the sheet into cyclohexane chairs with the presence of axially bound OH groups and 1, 3-ether [12, 13]; higher supersaturation of hydrogen that eliminate the ether groups and an enforced rigid linear structure [14]; the impact of fluorination [15]; and the development of oxygenation [16]. In 1998 the Lerf–Klinowski model was proposed [17, 18] which has since been depicted in most GO publications (and all of the reviews cited above) since this time. While instructive this model has a number of limitations, reflecting only the chemical connectivity and not the spatial distribution of the functional groups, and assuming aromatic ‘islands’ intermixed with oxidized aliphatic regions containing C–OH, epoxide groups and double bonds. No defects or rings with more or less than 6-members were included; nor were out-of-plane distortions disrupting the nearly flat structure beyond slight perturbations due to tetrahedrally bonded OH groups, or intrinsic ripples [19, 20]. More recently in 2006 Szabó et al [21] revised the Lerf–Klinowski model based on new evidence and reintroduced wrinkling via ribbons of cyclohexane chairs,
and the possibility of larger out-of-plane distortions. The Lerf–Klinowski model still persists however, even though it is largely inconsistent with the structures obtained by computational studies at all levels of sophistication, [22–25] and observations using electron microscopy [2, 21, 26, 27].

In this paper we revisit the structure of GO and determine the pure and truly representative structures for graphene nanoflakes using machine learning. We make no prior assumptions about the structure of GO, other than the range of concentrations of O and OH chemisorbing to the surfaces observed in experiments, and sample 20 396 random configurations at the electronic structure level. In all structures we observe the presence of ether, double bonds, aliphatic (cyclohexane) disruption, defects and significant out-of-plane distortions that depend on the size of the nanoflake. We identify 25 archetypal ‘pure’ GO structures which can be used to define the configuration space, capturing all of the complexity and diversity of the entire data set; and three prototypes that represent the averages of the three classes in the data set. Together these 28 structures, which are shown to be largely robust against changes in thermochemical conditions modeled using *ab initio* thermodynamics, can be downloaded and used together as a representative ‘little’ data set for future research, or individually when the profiles are matched to observations from conventional experiments.

2. Methods

2.1. Data set

To gather this data set a large number of virtual GO nanoflakes were systematically created to sample a wide range of sizes, topologies, oxygen concentrations, chemical groups and distributions. The data set contains 20 396 samples with areas ranging from 320 to 2457 Å² including four different main morphologies: hexagonal (49.5%), trigonal (14.3%), rectangular (30.5%), and rhombic (5.7%). The total number of atoms in each sample varies from 191 to 1949 including C, H, and O, and different ratios of armchair (AC) and zig zag (ZZ) edges were included. The density and distribution of oxygen groups have a significant role in deriving GO properties, so for each of the 24 primary pristine graphene nanoflakes numerous O/H concentrations were sampled, each with hundreds of random distributions (see figure S41 in the supporting information, which is available online at stacks.iop.org/NANOF/3/045001/mmedia). In each case the O/C ratio was between 4.05% and 52.08%, and the H/C ratio between 2.22% and 49.26%. In the majority of cases the edges were H-terminated (91.3%) with the remainder unpassivated to increase the structural diversity of the set. Figure 1 shows the distribution of ether and hydroxyl concentrations resulting from our sampling process. Note that the concentration of ether and hydroxyl is measured with respect to the total available oxygen, while the concentration of oxygen is measured with respect to the total number of atoms. A range of simple statistics are given in table 1 to provide an overview of the breadth of samples in our data set. The relaxed structure of each sample was determined at the electronic structure level using density functional tight binding (DFTB) method (as described in the supporting information). More than 30 million CPU hours were consumed in order to perform these calculations with this level of computational sophistication.

2.2. Feature extraction

Extracting a set of features that can well characterize the samples and explain the variance in our data set is a key factor to better understanding of the population of our ensemble. A comprehensive structural analysis was performed to extract 830 structural features. These features are grouped into four main categories:
thermodynamic (1 feature), structural (19 features), chemical (20 features), and statistical (790 features). The thermodynamic feature is the probability of observation, which was obtained using a Boltzmann distribution via the enthalpy of formation from the DFTB calculations and the thermodynamic treatment described below.

The 19 structural features characterize different geometrical aspects of the flake such as its average diameter, minimum and maximum dimensions, total area of flake, the anisotropy in the flake, and the ZZ and ACedge ratios. Attributes characterizing defects and out-of-plane distortions are included in this group. The 20 chemical features include the number and distribution of atomic and chemical species on each nanoflake (counts and concentration of each species, where concentration is counts normalized by total number of atoms) to provide useful overall information about each structure. Of the 790 statistical features, 16 characterize the atomic coordination of C, H and O; 268 characterize the ring statistics; 91 describes bond lengths between different species; 132 describe bond angles; 268 describe oxygen density distributions; 2 define particle and mass density; and 13 describe the distance distributions between different oxygen groups. The 281 features characterizing the oxygen distribution capture the differences between nanoflakes with similar graphene (primary) structures and oxygen concentrations, but different patterning of oxygen groups on the surface. For this purpose the following Gaussian probability (density) distribution function (PDF) is used:

$$\rho(x, y, z) = \sum_{i=1}^{\Omega_s} e^{-\alpha(r_i - r_{ref})^2},$$

where $r_{ref}$ is the reference location in space to evaluate the species-dependent probability distribution, $r_i$ is the distance of ith species of type $s$ from the reference location, and $\alpha$ is a smoothing parameter. Here, $\alpha$ is set such that the Gaussian function would fade at a distance $-4 \, \text{Å}$ (which is higher than van der Waals distance between the species on GO). This PDF can be used to define a local density at any point in space and for any type of species. However, these local densities are not normalized, which presents problems for nanoflakes of finite sizes. Lattice sites around the edges and corners of a nanoflake will have a different local environment than those in the interior; fewer carbons providing fewer opportunities for functionalization. To resolve this the local density of carbon atoms can be used to normalize each density distribution using:

$$\rho^\text{atom}(x, y, z) = \frac{\sum_{i=1}^{\Omega_s} e^{-\alpha(r_i - r_{ref})^2}}{\rho_C(x, y, z)},$$

where $\rho_C$ is the local density of carbon atoms at a reference location $(x, y, z)$. The density functions in equations (1) and (2) can provide local densities of carbon and oxygen atoms as well as ether and hydroxyl groups. The density functions are sampled at the location of each carbon atom, then different statistics such as mean value, standard deviation, skewness, and kurtosis are calculated in order to characterize the distribution. While, these statistics provide useful information on oxygen heterogeneity, they do not provide insights into

| Input feature           | Minimum | Maximum | Mean   | Standard deviation |
|------------------------|---------|---------|--------|--------------------|
| C atoms                | 150     | 1014    | 428.08 | 215.61             |
| O atoms                | 7       | 507     | 100.55 | 98.59              |
| H atoms                | 4       | 428     | 104.65 | 78.37              |
| Average diameter (Å)   | 20.24   | 56.10   | 34.44  | 11.09              |
| Anisotropy             | 1.17    | 2.67    | 1.60   | 0.45               |
| Total area (Å²)        | 320     | 2457    | 982.79 | 614.83             |
| Ether concentration (%)| 10      | 77      | 30     | 11                 |
| Hydroxyl concentration | 5       | 77      | 32     | 13                 |
| Oxygen concentration   | 4       | 37      | 24     | 5                  |

Table 1. Summarizing statistics for a number of important structural and chemical features of the ensemble of GO nanoflakes.
calculated. In each case \((x_i, y_i, z_i)\) is measured from the geometrical centroid of all carbon atoms as it closely represents the geometrical center of each nanoflake. The summation is over sampling points which is set to the location of carbon atoms, noting that the PDF is a continuous function and it can be sampled at any location. Sampling at the location of carbon atoms provides a good resolution of density and is intuitive. Additional features can also be determined by normalizing the inertia moments as follow:

\[
I_{iz}^{\text{norm}} = \frac{\sum_{i=1}^{n} (x_i^2 + y_i^2) \rho_i(x_i, y_i, z_i)}{\sum_{i=1}^{n} (x_i^2 + y_i^2)}.
\]

(5)

Similarly, the above normalization can be applied for all other inertia components.

2.4. Clustering

Cutler and Breiman, in which the algorithm alternates between consuming for large data sets. Therefore we have followed the alternating optimization procedure suggested by Cutler and Breiman, in which the algorithm alternates between finding the optimal \(\alpha\)'s for a given set of \(x\) mixtures and finding the optimal \(x\) mixtures for a given set of \(\alpha\)'s. The latter problem effectively optimizes \(\beta\) for a given value of \(\alpha\) [28]. In this iterative scheme each step requires the solution of several convex least squares problems of the following form: Find \(w_1, ..., w_p\) to minimize \(\| u - \sum_{k=1}^{q} w_k t_k \|_2^2\) for given \(u\) and \(t_1, ..., t_p\), subject to \(w_i \geq 0\) and \(\sum_{k=1}^{q} w_k = 1\). An in-house python package is developed to implement this algorithm and extract the archetypes for this study which as been made available online for download [29].

2.3. Archetypal analysis (AA)

AA is a matrix factorization method where an \(n \times m\) matrix \(X\) represents a multivariate data set with \(n\) observations and \(m\) variables; in our case \(n\) corresponds to the GO nanoflakes in the data set and \(m\) corresponds to the structural features [28]. AA aims to find an \(n \times m\) matrix \(Z\) that corresponds to the archetypal or ‘pure patterns’ in the data in such a way that each data point can be represented as a mixture of those archetypes. In other words, the AA yields the two \(n \times p\) coefficient matrices \(\alpha\) and \(\beta\) which minimize the residual sum of squares:

\[
\text{RSS} = \sum_{i=1}^{n} \| X_i - \sum_{k=1}^{p} \alpha_{ik} Z_k \|^2 = \sum_{i=1}^{n} \| X_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} X_j \|^2
\]

under the constraints:

(i) \(\sum_{k=1}^{p} \alpha_{ik} = 1\) with \(\alpha_{ik} \geq 0\) and \(i = 1, ..., n\).

(ii) \(\sum_{j=1}^{n} \beta_{kj} = 1\) with \(\beta_{kj} \geq 0\) and \(j = 1, ..., p\).

Constraint (i) indicates that predictors of \(X_i\) are finite mixtures of archetypes, \(X_i = \sum_{k=1}^{p} \alpha_{ik} Z_k\), while constraint (ii) implies that archetypes \(Z_j\) are convex combinations of the observations, \(Z_j = \sum_{j=1}^{n} \beta_{kj} X_j\). To solve (7) a general-purpose constrained nonlinear least squares algorithm can be used, but this could be time-consuming for large data sets. Therefore we have followed the alternating optimization procedure suggested by Cutler and Breiman, in which the algorithm alternates between finding the optimal \(\alpha\)'s for a given set of \(x\) mixtures and finding the optimal \(x\) mixtures for a given set of \(\alpha\)'s. The latter problem effectively optimizes \(\beta\) for a given value of \(\alpha\) [28]. In this iterative scheme each step requires the solution of several convex least squares problems of the following form: Find \(w_1, ..., w_p\) to minimize \(\| u - \sum_{k=1}^{q} w_k t_k \|_2^2\) for given \(u\) and \(t_1, ..., t_p\), subject to \(w_i \geq 0\) and \(\sum_{k=1}^{q} w_k = 1\). An in-house python package is developed to implement this algorithm and extract the archetypes for this study which as been made available online for download [29].

2.4. Clustering

Clustering methods are unsupervised pattern recognition techniques that group samples based on a similarity index. \(k\)-Means clustering which uses distance between the samples as a similarity index, is the most widely used technique due to its computational efficiency [30]. \(k\)-Means is a vector quantization method that partitions \(n\) observations (samples) into \(k\) clusters. Each observation belongs to the cluster with the nearest mean where the
mean of each cluster (centroid) is the prototype. The observations in each cluster can then be represented by its prototype structure. The number of clusters \( k \) must be defined \textit{a priori}, so that \( k \)-Means can group the \( n \) observations \( (x_1, x_2, \ldots, x_n) \) into \( k \leq n \) sets \( S = S_1, S_2, \ldots, S_k \) such that the within-cluster sum of squares is minimized:

$$
\text{arg} \min_{\mu} \left( \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \right),
$$

where a Euclidean distance is used as a measure of similarity and \( \mu_i \) is the mean of points in \( S_i \). The distance between the points is calculated in a \( d \)-dimensional space, where \( d \) is the number of variables (features).

A presumption in \( k \)-Means clustering is that the samples form clusters with approximately spherical shapes with similar size, which often yields poor results for dispersed and nonlinear data. Another drawback is in specifying the number of clusters \( k \); an inappropriate choice for \( k \) can also result in poor performance. While \( k \)-Means is one of the easiest clustering algorithm to apply, it can give misleading results as, given an input \( k \) it will cluster the data, even if there is no clusters present (null case).

Density-based clustering techniques overcome this limitation and are capable of identifying clusters of various sizes, shape and dispersivities. One of the most robust density-based clustering techniques is density-based spatial clustering of applications with noise (DBSCAN), which, unlike \( k \)-Means, does not require the number of clusters to be specified \textit{a priori}. DBSCAN does require two hyper-parameters, \( \epsilon \) and the minimum number of points to define dense region, which can significantly affect the quality of clustering [31, 32]. Another important issue of DBSCAN is in dealing with data sets with large variations in densities; setting parameters to high density can result in labeling clusters of low densities as noise, while a low density estimate may result in incorrect grouping of clusters.

Recently, a new clustering method has been introduced that has the advantage of including hyper-parameter optimization [33]. Iterative label spreading (ILS) is based on a general definition of a cluster and the quality of a clustering result, and is capable of predicting the number and type of clusters and outliers in advance of clustering, regardless of the complexity of the distribution of the data. ILS can be used to evaluate the results from other clustering algorithms, or perform clustering directly. It has been shown to be more reliable than alternative approaches for simple and challenging case (such as the null and chain cases) and to be ideal for studying noisy data with high dimensionality and high variance, as we have for our GO nanoflakes.

We have tested all three clustering techniques in this study.

### 2.5. Thermodynamics

The thermodynamic feature mentioned above (the probability of observation) can be determined using a Boltzmann distribution, by combining the results calculated from electronic structure simulations at the ground state, and the extensive thermochemical data measured at the standard state. Under the assumption that the change in electronic energies represents the most significant contribution to the change in free energy (i.e. other contributions cancel out), the free energy change in a reaction \( xA + yB + zC \rightarrow D \) at the standard state can be calculated as

$$
\Delta G^\circ(T) = \Delta_f G^\circ_B(T) - x \Delta_f G^\circ_A(T) - y \Delta_f G^\circ_B(T) - z \Delta_f G^\circ_C(T)
$$

$$
= [\Delta E_D(T_0) + \Delta \mu^\circ_D(T)] - x [\Delta E_A(T_0) + \Delta \mu^\circ_A(T)] - y [\Delta E_B(T_0) + \Delta \mu^\circ_B(T)] - z [\Delta E_C(T_0) + \Delta \mu^\circ_C(T)],
$$

where \( \Delta E(T_0) \) are the electronic energies with respect to the first principles reference state, \( \Delta \mu^\circ(T) \) are the chemical potentials of the reactants and products, and \( \Delta \mu^\circ \) are the Gibbs’ free energies of formation for the reactants and products which can be extracted from the thermochemical tables. If \( D \) is a solid \( \Delta \mu^\circ(D) \) it is typically orders of magnitude smaller than \( \Delta \mu^\circ(A) \), \( \Delta \mu^\circ(B) \) and \( \Delta \mu^\circ(C) \), and is usually omitted.

In this method, the gap between the ground state and the standard state in the chemical potential is filled by the connection energy

$$
\Delta \mu^\circ(T) = \mu^\circ(T) - \mu(T_0),
$$

where \( T_0 \) is 0 K, and \( \mu^\circ \) is the chemical potential under the standard pressure (\( P^0 = 1 \) atm). The chemical potential \( \mu^\circ(T) \) can be calculated using the heat capacities, \( C_p \), and entropy, \( S \), such that:

$$
\mu^\circ(T) = \mu^\circ(T_0) + \int_{T_0}^{T} C_p \,dT - T \int_{T_0}^{T} \frac{C_p}{T} \,dT - (T - T_0) S(T_0),
$$

where \( T_0 \) is the ambient room temperature (298.15 K), \( C_p \) and \( S(T_0) \) are obtained from the Shomate equation, equations:
\[
C_p = A + Bt + Ct^2 + Dt^3 + \frac{E}{T^2}
\]
\[
S(T_0) = A \ln t + Bt + \frac{C}{2}t^2 + \frac{D}{3}t^3 - \frac{E}{2t^2} + G
\] (13)
in which \(A\) to \(G\) are Shomate coefficients. Equation (12) can further be extended to include finite pressures, \(P\), using:
\[
\mu^0(T, P) = \mu^0(T) + RT \ln \left( \frac{P}{P_0} \right).
\] (14)
where \(P_0\) is the vacuum, and \(R\) is the gas constant. By calculating the chemical potential for each constituent element, the temperature- and pressure-dependent Gibb’s free energy can be obtained as:
\[
\Delta G = \left( E_{\text{tot}} - \sum_i N_i \mu_i^0(T, P) \right) / N_{\text{tot}}.
\] (15)

3. Results

3.1. Archetypes

To identify the pure structures of GO, AA was performed on the reduced space with a dimension of 224, including all 20,396 GO nanoflakes. The AA algorithm requires \textit{a priori} selection of the number of archetypes (similar to \(k\)-Means clustering) so we have used the elbow criteria as shown in figure 2(a). Here we can see that 25 archetypes can explain \(~70\%\) variance of our data set; effectively reducing the dimensions from 224 to 25. These 25 archetypal GO nanoflakes are the pure types (in 224-dimensions) and all other GO nanoflakes can in principle be described by a linear combination of these archetypes. This is demonstrated for our data set in figure 2(b), where we can see the simplex plot of all the GO structures in this 25-dimensional archetype-space. We can also see that although we have a large number of GO nanoflakes in the set, this set is by no means exhaustive, as there are regions of the archetype-space that are not populated.

In addition to significantly reducing the dimensionality by transforming our feature-space into archetype-space, AA is capable of identifying archetypal structures that were not present in the original set, which can be characterized (in the 224 dimensional feature-space) to gain some insight into unseen data. By employing an empirical cumulative distribution function (ECDF), the histogram profiles of the archetypes are extracted and included for all 25 archetypes are provided in the supporting information. Due to the large number of dimensions in the feature-space each profile is split into four histograms. The histograms reveal how influential each feature is in deciphering the identity of the archetype; each archetype is high in some features and low in others. The archetypes are also ordered by ranking them based on the \(\sum_{i=1}^{n} || X_i - \alpha_k Z_k ||^2 \) value, which is the summation of errors in approximating data points with an individual archetype.

As mentioned above, archetypes do not need to be in the original set to exist and to be identified, but in a sufficiently large set it is possible that they will be present, or that some very close matches have been captured. In our case we calculated the Euclidean distance between each archetype and the original data points in the feature-space, and used the ECDF to extract the profile of these best matching nanoflake. The profiles of the best matching nanoflake is overlaid with each archetype in the supporting information to quantify the quality of the match, which is captured more succinctly in the radar plot in figure 2(c). This plot indicates how close each matched GO nanoflake is to the corresponding archetype; highlighting that a large population of similar structure does not always guarantee finding a pure type \textit{a priori}. A20 is present in our data set, but this region of the archetype-space is sparsely populated. The region around A8 is heavily populated, but the best match to A8 is not a perfect match. It would take hundreds of thousands (possibly millions) of samples to spontaneously capture the pure types without AA to identify what their profiles look like, but once identified these small collections of matched structures have been shown to be invaluable in summarizing the properties of larger data sets [34] or providing a guide as to which structures are worthy of further investigation [35, 36]. These best matching structures for each archetype are illustrated in figure 3, and are available for download [37].

To provide some straightforward interpretation of the diversity of the archetypal GO nanoflakes table 2 lists a set of selected features taken from the detailed profiles. Here we can see that the archetypes span a wide range of sizes, with vastly different concentrations of hydroxyl, ether and ZZ edges (inputs into the DFTB+ simulations), and point defects or planar distortion (outputs from the DFTB+ simulations). The defect concentration appears universally low, which reflects the fact that few defects (broken C–C bonds) formed during the DFTB+ relaxation, but even there the entire range is represented. In general there is no underlying pattern (such as smaller nanoflakes having higher ether concentration, etc) since these are by definition the extreme cases around the convex hull. The only exception to this observation is that larger nanoflakes exhibit larger distortions, which
is simply due to larger nanoflakes being able to accommodate distortions with a magnitude or period that is larger than the smaller nanoflakes in size.

Beyond providing specific information on the nature of the pure types of GO nanoflakes, AA is also a powerful tool for mapping the high-dimensional feature-space to 2-dimensions, and like other visually interpretable dimension reduction methods high-dimensional patterns in the data set can be seen by color coding each point with additional information [38]. Figure 4 presents the same simplex plot from figure 2(b) encoded by the size, shape and composition of each sample. In each case the shape of the symbol reflects the shape of the nanoflake (hexagonal, rectangular, rhombic or trigonal), and the relative size of each symbol reflects the relative size of the nanoflakes. Figures 4(a) and (b) are colored by the ether and hydroxyl concentrations, respectively. Figures 4(c) and (d) are colored by defects concentration and distortion index. While these plots are highly complex (and a number of sample are mapped into similar regions in the convex hull, obscuring the detail) it is easy to see distinct groupings where structures with high similarity in 224 dimensions are adjacent in 25-dimensional archetype space, and to understand why the archetypes are important structures. For the purposes of clarity these plots are separated by shape and reproduced in the supporting information. Color encoding of this type is not limited to the features used to extract the archetypes; any of the 830 original features, or any future property label, can also be visualized in this way.

3.2. Prototypes

While archetypes are an excellent way of defining the structure space and describing all possible structures in it, they cannot average over the sample as a whole. Prototypes are the representative data points, defined by the centroids of clusters with similar characteristics in 224 dimensions, and represent the average nanoflakes in the data set. To identify the prototypes we tested three different clustering techniques (k-Means, DBSCAN and ILS), with ILS providing superior performance. Both k-Means and DBSCAN suffer from some limitations that are
challenged by our highly nonlinear and imbalanced data set, and require subjective hyper-parameters to be estimated a priori (i.e. the number of clusters, epsilon and minimum number of samples, respectively). The elbow criteria and a grid search technique was used to optimize these parameters, but the simplicity and advantages of ILS verifiably identified our three clusters (which k-Means did not) without mistaking the noise (as DBSCAN did). ILS clustering does not require a priori hyper-parameter estimation or optimization, but must be initialized so the results were repeated with various starting points and always returned same three clusters. The results for the ILS clustering, and the identity of the centroid prototypes are described below (along with the profile histograms in supporting information), and the corresponding results for k-Means and DBSCAN are provided in the supporting information for completeness.

To provide insight into how the samples are clustered a simplex plot of data set is shown in figure 5 where each sample is colored by the ILS cluster, in the low dimensional archetype-space. The prototypes are highlighted by the crosses, and are separated individually in radar plots to show the contribution of the 25 archetypes to the description of the three prototypes identified using ILS (figures 3(b)–(d)). The largest cluster (shown in red) contains 77.8% of the samples, while the other two include 15.8% (green) and 6.4% (purple) of the nanoflakes, respectively. This is similar to the results for DBSCAN (shown in the supporting information), which also recognized a large cluster containing 76.5% of the data (with two more clusters with 5.1% and 1.4%, and four tiny clusters containing only a few points) but interpreted 15.0% of data as noise. k-Means clustering was performed with five clusters (see supporting information) resulting in clusters with population of 36.5%, 40.4%, 7.2%, 9.3%, and 6.6%, which reflects its pre-assumption of a spherical distribution of 5 enforced clusters. As shown in the supporting information, k-Means splits the large cluster into two smaller clusters for no
misleading results. Using the archetypes and prototypes ensures that the most inclusive and representative
turn these advantages into disadvantages. Incorrect assumptions about the concentration of oxygen or oxygen-
diffusion. However, the indirect nature of the interventions make tuning GO challenging, since the structure is
apparent reason. This cluster-encoded simplex plot has been separated by shape and re-plotted in figure S37 in
the supporting information for clarity.

To understand the characteristics of the ILS prototypes the histogram profiles are extracted and compared
with their closest matching nanoflake, as shown in figures S38, S39 and S40, and a number of selected features
are listed in table 3 (for both the prototypes and the closest matching nanoflakes). In contrast to the archetypes,
all of the prototypes differ mainly in their total area, and the degree of out-of-plane distortion (which is related to
the area, as mentioned above), which can be confirmed by visual inspection of the marker sizes in figure 5(a).
This trend is remarkably consistent amongst the clustering techniques (see supporting information), even when
clusters have been mistaken (DBSCAN and k-Means). As confirmation that the prototypes are averages of the
nanoflakes in each cluster, we generated radar plots of the best matching prototypical nanoflakes in the 25-
dimensional archetype-space (figures 5(b)–(d)), where we can see that they are centrally located near the origin
and require numerous archetypes to define them.

The atomic configurations of the three ILS prototypes are shown in figure 6 and provided along with the
archetypes for download [37].

Table 2. Selected features of the 25 archetype, with the corresponding values for the best matching nanoflakes in the data set provided in braces.

| Archetype | Anisotropy | Area, Å² | Ether, % | Hydroxyl, % | ZZ, % | Defect, % | Distortion, Å |
|-----------|------------|----------|---------|-------------|-------|-----------|--------------|
| A1        | 1.2 (1.2)  | 2439 (2426) | 28.1 (28.4) | 57.3 (56.3) | 40.4 (3.4) | 3.1 (3.2) | 15.2 (12.4) |
| A2        | 1.2 (1.2)  | 382 (382) | 24.3 (24.1) | 74.8 (74.7) | 9.1 (9.1) | 4.7 (4.4) | 8.4 (11.5) |
| A3        | 1.2 (1.2)  | 320 (320) | 47.7 (50.0) | 20.4 (18.1) | 77.8 (77.8) | 1.6 (2.7) | 7.5 (7.4) |
| A4        | 1.9 (1.9)  | 1330 (1330) | 54.3 (54.5) | 27.6 (27.3) | 44.0 (44.0) | 1.5 (1.9) | 15.8 (14.2) |
| A5        | 1.2 (1.2)  | 329 (329) | 57.8 (54.6) | 15.4 (14.4) | 68.1 (77.8) | 3.2 (4.0) | 8.2 (6.6) |
| A6        | 1.2 (1.2)  | 1734 (1734) | 17.4 (17.4) | 34.8 (34.8) | 90.5 (90.5) | 0.3 (0.3) | 7.8 (5.7) |
| A7        | 1.2 (1.2)  | 382 (382) | 66.9 (66.1) | 32.1 (32.2) | 9.1 (9.1) | 6.2 (4.4) | 11.0 (11.6) |
| A8        | 1.3 (1.2)  | 324 (320) | 29.7 (35.5) | 37.5 (32.3) | 70.0 (77.8) | 3.5 (2.7) | 6.9 (5.4) |
| A9        | 1.2 (1.2)  | 1719 (1734) | 34.7 (34.8) | 17.4 (17.4) | 88.4 (90.5) | 0.8 (0.8) | 8.6 (8.0) |
| A10       | 1.2 (1.2)  | 325 (320) | 39.9 (40.0) | 29.4 (30.0) | 72.6 (77.8) | 3.0 (4.0) | 7.6 (8.3) |
| A11       | 1.9 (1.8)  | 558 (503) | 12.2 (12.2) | 35.7 (36.6) | 39.1 (38.5) | 0.0 (0.0) | 6.0 (6.7) |
| A12       | 2.0 (2.1)  | 502 (503) | 36.6 (36.6) | 12.1 (12.2) | 54.3 (59.0) | 1.5 (1.7) | 8.3 (7.5) |
| A13       | 2.1 (2.0)  | 713 (995) | 39.2 (42.9) | 13.5 (14.3) | 81.4 (94.7) | 7.9 (6.8) | 13.3 (15.8) |
| A14       | 1.6 (1.2)  | 1351 (1263) | 41.0 (48.4) | 24.3 (16.1) | 30.3 (4.8) | 1.9 (1.5) | 13.4 (13.5) |
| A15       | 1.2 (1.2)  | 412 (320) | 20.3 (19.0) | 11.9 (9.5) | 65.6 (77.8) | 0.0 (0.0) | 5.2 (3.5) |
| A16       | 1.7 (1.6)  | 2057 (2096) | 36.9 (35.7) | 31.3 (35.0) | 57.0 (49.4) | 2.2 (1.6) | 18.1 (13.7) |
| A17       | 2.2 (2.0)  | 873 (995) | 13.0 (14.5) | 39.2 (42.0) | 81.6 (94.7) | 0.7 (0.9) | 8.9 (14.2) |
| A18       | 1.2 (1.2)  | 320 (320) | 28.5 (31.7) | 31.9 (25.6) | 77.8 (77.8) | 1.5 (1.3) | 6.8 (6.1) |
| A19       | 1.2 (1.2)  | 334 (320) | 50.6 (48.0) | 24.5 (22.0) | 62.7 (77.8) | 4.0 (1.3) | 9.7 (7.4) |
| A20       | 1.8 (1.6)  | 1991 (2096) | 34.4 (34.9) | 33.5 (35.9) | 27.6 (49.4) | 2.0 (1.6) | 20.2 (22.3) |
| A21       | 1.5 (1.6)  | 386 (335) | 25.7 (21.3) | 39.1 (42.6) | 48.6 (48.4) | 2.8 (2.5) | 10.4 (8.1) |
| A22       | 1.6 (1.6)  | 347 (335) | 15.4 (15.6) | 31.6 (31.3) | 47.7 (48.4) | 4.0 (3.8) | 7.8 (6.0) |
| A23       | 1.2 (1.2)  | 382 (382) | 49.9 (50.0) | 49.9 (50.0) | 9.1 (9.1) | 1.6 (1.1) | 8.5 (8.1) |
| A24       | 2.7 (2.7)  | 671 (671) | 11.4 (11.4) | 31.8 (31.8) | 66.0 (66.0) | 1.3 (1.3) | 7.0 (7.0) |
| A25       | 1.4 (1.6)  | 347 (33) | 26.1 (19.4) | 37.5 (44.1) | 56.1 (48.4) | 1.9 (1.3) | 8.4 (7.9) |

4. Discussion

One of the advantages of using GO in various applications is the ability to tune the properties indirectly via
structural interventions; for example, GO can be a semiconductor or an insulator depending on the degree of
oxidation. However, the indirect nature of the interventions make tuning GO challenging, since the structure is
difficult to characterize due to its nonstoichiometric composition that depends on the thermochemical
conditions, its strongly hygroscopic nature, and its slow decomposition above 60 °C–80 °C. A lack of control can
turn these advantages into disadvantages. Incorrect assumptions about the concentration of oxygen or oxygen-
containing groups, the distribution of oxygen on the surface, the location and nature of defects, and the degree of
out-of-plane distortion can lead to inappropriate data sets being generated and used, which ultimately lead to
misleading results. Using the archetypes and prototypes ensures that the most inclusive and representative
structures are assumed, and dramatically increases efficiency. A computational study using the archetypes
exclusively will capture all of the complexity and diversity inherent in our 20 396 nanoflake data set, effectively
reducing the computational cost by three orders of magnitude (99.87%). In contrast, a computational study
using the prototypes exclusively will provide an average of all of the 20 396 nanoflake data set, effectively reducing the computational cost by four orders of magnitude (99.98%).

One question that arises however, is how these results relate to the low energy nanoflakes, which are often incorrectly assumed to be ‘representative’ of polydispersed samples? The low energy structure, or stereotypes, are generally special cases and exist as outliers, but it is interesting to see where they reside in the 25-dimensional archetype space with respect to the prototypes. In figure 7(a) we provide a simplex plot showing the location of the 77 stereotypes in archetype-space, colored by the total electronic energy per atom. We can see from this plot that the low energy stereotypes are not distributed across the space, and are therefore not representative of the data set; instead being group in the vicinity of A5. We can also see that they are all hexagonal, lacking the diversity of the archetypes; and relatively small, lacking the diversity of the prototypes. In figure 7(b) we can see the linear contribution of the archetypes needed to describe the lowest energy nanoflake, and note the significant imbalance. Using low energy stereotypes to represent the data set would clearly be misleading.

Another question is how robust the archetypes and prototypes are with respect to changes in the data set? Firstly, AA assumes that the data set remains unchanged, and removal of samples will impact the outcome. The archetypes for one set will be different from the archetypes of another set, which is why we have used over 20 000 samples covering such a large range of configurations here. Secondly, AA uses a random seed and so it is best to use as recursive algorithm to test random seeds until consistent set of archetypes is obtained [36]. Previous work has shown that the profiles of archetypes is sensitive to the type of distribution in the data set [39], but remarkably insensitive to the size range [36]. This leaves the remaining issue of how robust the archetypes are with respect to experimentally relevant conditions, such as changes in temperature, pressure or the surrounding environment.

To extend the ground state energy to finite temperatures we have used a first principles thermodynamics method [40], with energies obtained using the original electronic structure simulations (temperature $T \approx 0$ K, pressure $P = 0$ Pa), and calculated thermodynamic probabilities of nanoflakes at different temperatures and

Figure 4. Simplex plot of the data set in the 25-dimensional archetype-space, encoded by (a) the ether concentration (Range: 10.0%–77.2%), (b) the hydroxyl concentration (Range: 5.0%–77.2%), (c) the defect concentration (Range: 0.0%–16.7%), and (d) the mean absolute error of the out-of-plane distortion (Range: 1.8 Å–30.7 Å).
pressure. An advantage of this method is the ability to estimate the probability under different environmental conditions, such as aerobic or anaerobic, humid or dry, by adjusting the oxygen supersaturation or the thermodynamic reference state. It is also possible to combine results (probabilities) weighted by competing reference states to predict the impact of (for example) 40% humidity. The thermodynamic probability can also be expanded to three (or more) extra features to study how the identity of the archetypes or prototypes are influenced by different processing or storage conditions; how pure, or representative, they really are.

To investigate the effect of environmental conditions such as temperature, pressure, and degree of ambient humidity, the thermodynamic probabilities were calculated using a Boltzmann distribution and the Gibb’s free energy (as described above, using the QuickThermo software [41]). These three parameters were varied in intervals ($\delta$) in a range as follows: $T = \{300 \text{ K}, \ldots, 750 \text{ K}\}$; $\delta T = 50 \text{ K}$; $P = \{100 \text{ kPa}, \ldots, 200 \text{ kPa}\}$; $\delta P = 50 \text{ kPa}$, and humidity, $H = \{0\%, \ldots, 100\%; \delta H = 20\%\}$. For each combination of $(T, P, H)$ the

![Figure 5](image)

**Figure 5.** (a) Simplex plot of all samples colored by clusters identified using ILS clustering. The cross points are the prototypes of each cluster mapped on convex hull of archetypes. (b)–(d) Radar plots showing the contribution of the 25 archetypes in describing the three prototypes, P1, P2 and P3, respectively. The simplex plots separated by shapes, and profile histograms of each prototype, are provided in the supporting information.

| Prototypes | Anisotropy | Area, Å² | Ether, % | Hydroxyl, % | ZZ, % | Defect, % | Distortion, Å |
|------------|------------|----------|---------|-------------|------|-----------|--------------|
| P1         | 1.5 (1.2)  | 811 (697)| 27.7 (22.7)| 32.1 (22.7)| 38.5 (6.7)| 1.7 (1.3) | 8.4 (8.1)    |
| P2         | 1.7 (1.2)  | 967 (873)| 31.8 (33.3)| 28.8 (33.3)| 58.4 (86.7)| 2.4 (1.6) | 11.0 (11.3)  |
| P3         | 1.4 (1.1)  | 1954 (2457)| 33.8 (28.2)| 49.1 (38.2)| 38.6 (92.0)| 2.8 (2.4) | 15.1 (17.1)  |

Table 3. Selected features of the prototypes identified using ILS.
thermodynamic probability was calculated for all samples in our data set and AA repeated. To track the \((T, P, H)\)-dependent changes \((T, P, H) = (300 \text{ K}, 100 \text{ kPa}, 0\%)\) was selected as the baseline, and each subsequent archetype mapped to the convex hull of the baseline set. Figure 8 presents the simplex plot of the changes in the archetypes with respect to variations in \(T, P\) and \(H\), where each point is encoded by the temperature or humidity level as defined in figure 8(b). As we can see, the identity of the majority of the archetypes remain unaffected by changes in the environmental conditions, remaining on the vertices of the convex hull, regardless of \(T\) or \(H\). A1, A2, and A7 however, experience a sudden and considerable displacement at higher temperatures, with A8 and A18 also exhibiting a temperature-dependent displacement. These archetypes describe regions of data set which are sparsely populated and experience high variance. In all cases changing the pressure in the specified range did not affect the archetypes.

The same procedure was used to track changes in the prototypes and stereotypes, using only the extreme conditions, and it was found that they are unaffected by the changes in the environmental conditions. This was expected since the average (over all results) and the lowest energy scale with temperature, pressure or humidity.

5. Conclusions

The structure of graphene oxide is critical to its performance in a variety of applications, and so the configuration of model structures is also critical to predictions made in advance of synthesis and characterization. Using a large and diverse data set of over 20 000 GO nanoflakes fully relaxed at the electronic structure level, characterized by an exhaustive list of 830 chemical, structural and topological features we have isolated the pure and representative form of GO using machine learning. We find that the 830 dimensional space can be reduced to 224 dimensions, and used in conjunction with multivariate AA to identify 25 archetypal structures, and clustering analysis to identify 3 prototypes that can represent the entire ensemble set. The archetypes and
prototypes bear similarity to the Lerf–Klinowski model, but include much more structural complexity (such as defects and distortions) and capture the importance of distributions of oxygen groups, in addition to the type and concentrations. They cover the entire size range of the ensemble, with different ether, hydroxyl and defect concentrations, and different oxygen distributions. By applying ab initio thermodynamics we show that these representative GO nanoflakes are largely robust against changes in the temperature, pressure and supersaturation of water in the surrounding environment. These 28 GO nanoflakes can be used collectively as a small data set with a fraction of the computational cost in future work, or independently as an exemplar of graphene oxide with the required oxidation to match experiments.

Acknowledgments

Computational resources for this project were supplied by the National Computing Infrastructure national facility under grant q27. AB also acknowledges Lin Lai.

ORCID iDs

Amanda J Parker https://orcid.org/0000-0003-2207-744X
Amanda S Barnard https://orcid.org/0000-0002-4784-2382

References

[1] Stankovich S, Dikin D A, Piner R D, Kohlhaas K A, Kleinhammes A, Jia Y, Wu Y, Nguyen S T and Ruoff R S 2007 Carbon 45 1558
[2] Zhu Y, Murali S, Cai W, Li X, Suk J W, Potts J R and Ruoff R S 2010 Adv. Mater. 22 3906
[3] Perrozzi F, Prezioso S and Ottaviano L 2014 J. Phys.: Condens. Matter 27 013002
[4] Dimiev A M and Tour J M 2014 ACS Nano 8 3060
[5] Chen D, Feng H and Li J 2012 Chem. Rev. 112 6027
[6] Shareena T P D, McShan D, Dasamahapatra A K and Tchounwou P B 2018 Nano Micro Lett. 10 53
[7] Rowley-Neale S J, Randvir E P, Dena A S and Banks C E 2018 Appl. Mater. Today 10 218
[8] Ahmad H, Fan M and Hui D 2018 Composites B 145 270
[9] Li D, Muller M B, Gilje S, Kaner R B and Wallace G G 2008 Nat. Nanotechnol. 3 101
[10] Li F, Jiang X, Zhao J and Zhang S 2015 Nano Energy 16 488
[11] Hofmann U and Holst R 1939 Ber. Dtsch. Chem. Ges. A/B 72 754
[12] Ruess G 1946 Monatsh. Chem. 76 381
[13] Mermoux M, Chabre Y and Rousseau A 1991 Carbon 29 469
[14] Scholz W and Boehm H P 1969 Anorg. Allg. Chem. 369 327
[15] Nakajima T and Matsuo Y 1994 Carbon 32 469
[16] Hontoria-Lucas C, López-Peinal A J, López-González I D, Rojas-Cervantes M I and Martín-Aranda R M 1995 Carbon 33 1585
[17] He H, Klinowski J, Forster M and Lerf A 1998 Chem. Phys. Lett. 287 53
[18] Lerf A, He H, Forster M and Klinowski J 1998 J. Phys. Chem. B 102 4477
[19] Fasolino A, Los J H and Katsnelson M I 2007 Nat. Mater. 6 858
[20] Barnard A S and Snook I K 2012 Nanoscale 4 1167
[21] Szabó T, Berkesi O, Forgó P, Josepovits K, Sanakis Y, Petridis D and Dékány I 2006 Chem. Mater. 18 2740
[22] Bagri A, Mattevi C, Acik M, Chabal Y J, Chhowalla M and Shenoy V B 2010 Nat. Chem. 2 581

Figure 8. Simplex plot of archetype sets at different temperatures, pressures, and humidities mapped on the convex hull of the set from $T = 300 \text{ K}$, $P = 100 \text{ kPa}$, $H = 0\%$.  

![Figure 8. Simplex plot of archetype sets at different temperatures, pressures, and humidities mapped on the convex hull of the set from $T = 300 \text{ K}$, $P = 100 \text{ kPa}$, $H = 0\%$.](image-url)
[23] Shen X, Lin X, Yousef N, Jia I and Kim J K 2014 Carbon 66 84
[24] Rozada R, Paredes J I, López M J, Villar-Rodil S, Cabria I, Alonso J A, Martínez-Alonso A and Tascón J M D 2015 Nanoscale 7 2374
[25] Chaban V V and Prexhbo O V 2015 Nanoscale 7 17055
[26] Mkhoyan K A, Contryman A W, Silcox J, Stewart D A, Eda G, Mattevi C, Miller S and Chhowalla M 2009 Nano Lett. 9 1058
[27] Gómez-Navarro C, Meyer J C, Sundaram R S, Chuvilin A, Kurasch S, Burghard M, Kern K and Kaiser U 2010 Nano Lett. 10 1144
[28] Cutler A and Breiman L 1994 Technometrics 36 338
[29] Motevalli Soumehsaraei B and Barnard A 2019 Archetypal Analysis Package, v1. CSIRO Software Collection (https://doi.org/10.25919/5d3958889f7ff)
[30] Anderberg M B 1973 Cluster Analysis for Applications, Probability and Mathematical Statistics 1st edn (New York: Academic)
[31] Ester M, Kriegel H P, Sander J and Xu X 1996 Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-1996) (AAAI Press) p 226
[32] Sander J, Ester M, Kriegel H P and Xu X 1998 Data Mining and Knowledge Discovery vol 2 (Berlin: Springer) p 169
[33] Parker A J and Barnard A S 2019 Adv. Theo. Sim. (https://doi.org/10.1002/adts.201900145)
[34] Fernandez M and Barnard A S 2015 ACS Nano 9 11980
[35] Swann E, Fernandez M, Coote M L and Barnard A S 2017 ACS Comb. Sci. 19 544
[36] Yan T, Sun B and Barnard A S 2018 Nanoscale 10 21818
[37] Motevalli Soumehsaraei Band Barnard A 2019 Graphene Oxide Nanoflake Archetypes and Prototypes, v1. CSIRO Data Collection (https://doi.org/10.25919/5d1304132364a)
[38] Sun B and Barnard A S 2019 J. Phys.: Mater. 2034003
[39] Fernandez M, Wilson H F and Barnard A S 2017 Nanoscale 9 832
[40] Zhang W, Smith J R and Wang X G 2004 Phys. Rev. B 70 024103
[41] Motevalli Soumehsaraei B and Barnard A 2019 QuickThermo, v1. CSIRO Software Collection (https://doi.org/10.25919/5d39589c523d1)