ACS Combinatorial Science

Letter

pubs.acs.org/acscombsci

# Machine Learning Prediction of the Energy Gap of Graphene Nanoflakes Using Topological Autocorrelation Vectors

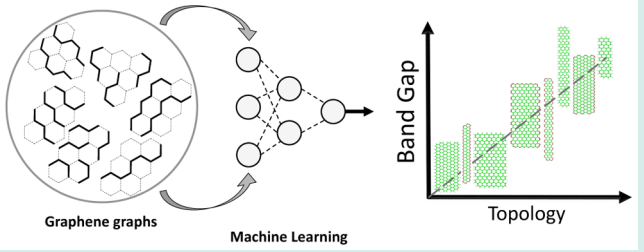
Michael Fernandez,*[,†] Jose I. Abreu,[‡] Hongqing Shi,[†] and Amanda S. Barnard[†]

[†]Data61, CSIRO, 343 Royal Parade, Parkville, Victoria 3052, Australia

[‡]Departamento de Ingeniería Informática, Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Alonso de Ribera 2850, Concepción, Chile

Supporting Information

**ABSTRACT:** The possibility of band gap engineering in graphene opens countless new opportunities for application in nanoelectronics. In this work, the energy gaps of 622 computationally optimized graphene nanoflakes were mapped to topological autocorrelation vectors using machine learning techniques. Machine learning modeling revealed that the most relevant correlations appear at topological distances in the range of 1 to 42 with prediction accuracy higher than 80%. The data-driven model can statistically discriminate between graphene nanoflakes with different energy gaps on the basis of their molecular topology.




**KEYWORDS:** *graphene, band gap engineering, machine learning, molecular topology, topological autocorrelation vectors*

The unique electric,[1] magnetic, and optical properties[2] of graphene, single atomic layers from graphite, have attracted considerable attention in recent years. By tuning of graphene properties, which are intrinsically related to its shape, size, and edge conformation, graphene sheets could be incorporated into a wide variety of electronics, optoelectronics, and electromagnetic devices. However, controlling the precise structure of individual graphenes remains challenging.[3] To circumvent the need for exquisite control at the atomic level,[4] computer simulations can elucidate how the nanostructure variability affects the functional properties.

The fact that graphene lacks a band gap around the Fermi level, which is the defining feature of semiconductor materials, limits the control of the conductivity of graphene by electronic means, hampering its revolutionary application in microelectronics. Theoretically, the band gap can be tuned by periodic modulations of the graphene lattice[5] and control of the zigzag and armchair edges.[6] Meanwhile, experimental evidence has been reported for band gap opening by patterned adsorption of atomic hydrogen onto the Moiré superlattice positions[7] and manipulation of the width of graphene nanoribbons and hydrogen passivation on the edges.[8] Alternatively, assembly of two-dimensional atomic crystals into stacks, i.e., graphene/boron nitride heterostructures, has emerged as a very promising system for band engineering of graphene.[9]

Experimental and theoretical evidence shows that differences in the electronic structures and associated properties of nanomaterials can be linked to discrepancies in physical structure, which can be measured using structural fingerprints or molecular descriptors. These fingerprints encoding topological, geometrical, or/and electronic features[10] can be conveniently combined with machine learning techniques to unravel complex structure−property patterns and build accurate predictive models.[10] Recently, we found that the energy gap ($E_G$) between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of graphene nanoflakes can be predicted from the interatomic distance distribution[10] and geometrical features[11] with accuracies of 70% and 90%, respectively. However, there is still room for further exploration of how the topology of the molecular graph can impact the energy gap of graphene nanoflakes.

In this Letter, we demonstrate that the topology of the molecular graph can accurately predict the energy gap of graphene nanoflakes regardless of atom position information, which could thereby accelerate the efficient rational design of functional graphene nanomaterials. More specifically, machine learning models have been developed to correlate graph topological fingerprints to the energy gap of graphene nanoflakes in quantitative terms. A virtual data set of 417 structures was used to calibrate correlation models with topological autocorrelation scores derived from graphene molecular graphs to predict the $E_G$ of another set of 205 nanoflakes as a parametric function of the molecular graph topology. In brief, the remainder of this Letter is organized as follows: first, a description of the data set of graphene nanoflake structures and the topological autocorrelation descriptors used

**Received:** June 30, 2016
**Revised:** August 23, 2016

DOI: 10.1021/acscombsci.6b00094
*ACS Comb. Sci.* XXXX, XXX, XXX−XXX

to calibrate the machine learning models is given, and then an account of how the predictive models were trained and validated is provided along with concluding remarks.

The influence of the graphene nanoflake topology on the size of the energy gap was explored using a data set of 622 virtual nanographene samples with a large range of sizes (16 to 2176 carbon atoms) that includes the ranges observed experimentally, which were simulated using the density functional tight binding (DFTB) method described elsewhere.[12] The set of structures is available free of charge at the CSIRO Data Access Portal (DAP).[13]

The set of topological autocorrelation scores (ATS) described below is a very convenient way to describe the topology of graphene nanoflakes, as they are independent of the original atom numbering, so they are canonical and independent of the size of the molecular graph. Therefore, a substantial reduction in data by limiting the topological distance can be achieved, which allows the analysis of a large data set with a wide range of molecular size and chemical diversity. ATS and related descriptors were first proposed to relate the molecular graphs of organic compounds to their biological activities[14] and later successfully extended to elucidate how the distribution of physicochemical features along protein sequences affects their conformational stability[15] and ligand affinities.[16]

In Figure 1, the graphene nanoflake is represented as a molecular graph, which can be transformed into a numerical ATS code that is independent of the size of the structure. To differentiate between unpassivated and passivated edges, ATS values are weighted using bond order values of 2 and 3, respectively, while internal fully connected carbon atoms are assigned bond order values of 3. The ATS vectors can be interpreted as correlations of the bond order at different topological distances $L$ in the molecular graph according to eq 1:

$$\mathrm{ATS}^L = \sum_{ij}^{N} \delta_{ij}^L \times P_i \times P_j \tag{1}$$

where $P_i$ and $P_j$ are the values of a property of the graph nodes $i$ and $j$, in this case the bond order of the carbon atoms in graphene, and $\delta_{ij}^L$ is a delta function defined as follows:

$$\delta_{ij}^L = \begin{cases} 1 & \text{if } L = d_{ij} \\ 0 & \text{otherwise} \end{cases}$$

In order to simplify the autocorrelation indices, the summation in eq 1 runs over only the $N$ pairs of carbon atoms that are not fully connected in the molecular graph, i.e., edges and defects, while $d_{ij}$ is the topological distance or shortest path between carbon atoms $i$ and $j$ including all connections in the molecular graph, i.e., edges, defects, and internal connections, in the range from 1 to 100. The ATS vectors computed by eq 1 are not normalized by the number of atom pairs, so they can easily discriminate among different structure sizes.

In the case of graphene nanoflakes, the use of topological indices to characterize the molecular structure is particularly convenient considering that the edge topology and overall shape have been shown to influence the energy gap to a large extent.[6] In this respect, connectivity information in graphene can be derived from experimental structure analysis of graphene films using high-resolution transmission electron microscopy (HRTEM),[17] as has been successfully reported by McNerny et al.[18]

The ATS fingerprints averaged over hexagonal, rectangular, and trigonal graphene nanoflakes in Figure 2a illustrate differences among the graph representations for three shapes, which mainly differ in height, with rectangular nanoflakes showing the higher values, followed by trigonal and hexagonal structures in that order. Rectangular and trigonal ATS vectors show a maximum around topological distance of 20 steps, while hexagonal structures exhibit a maximum at around 30

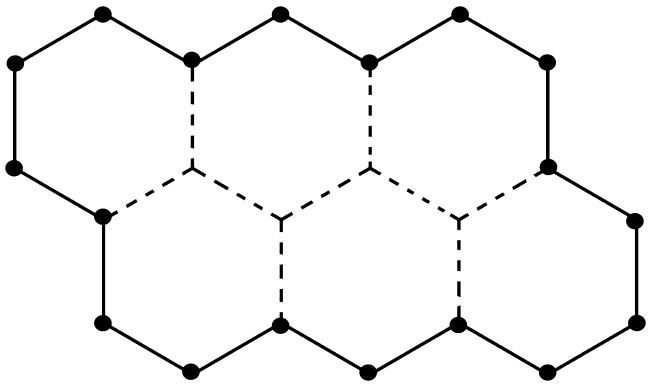


**Figure 1.** Molecular graph representation of a graphene nanoflake with edges connected by solid lines and all other interior nodes connected by dashed lines.

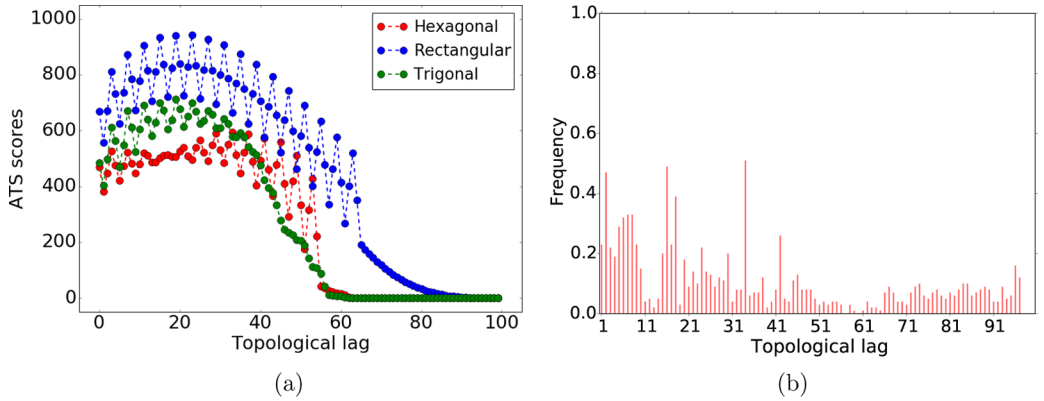



**Figure 2.** (a) ATS averaged over hexagonal, rectangular, and trigonal graphene nanoflakes. (b) Histograms of the ATS in the optimum SVM models of the energy gap of graphene nanoflakes in 200 independent GA runs.



DOI: 10.1021/acscombsci.6b00094
ACS Comb. Sci. XXXX, XXX, XXX−XXX

topological steps in the molecular graphs. However, the complexity of the ATS profiles suggests that sophisticated pattern recognition techniques rather than simple correlation methods can correlate the $E_G$ values with a selected number of relevant ATS.

Simple regression models, namely, multiple linear regression (MLR) and binary decision tree (DT) calibrated on a training set of 70% of the data set, yielded cross-validation correlation coefficients of ∼0.73 for the entire set of autocorrelation vectors (see the Supporting Information for details). Meanwhile, more sophisticated nonlinear mapping techniques, namely, support vector machines (SVM) and artificial neural networks (ANN), yielded cross-validation correlation coefficients of ∼0.77 (see the Supporting Information for details on the machine learning models).

We further improved the machine learning models by exploring optimum combinations of ATS vectors simultaneously that evolve for different generations according to a genetic algorithm (GA) previously described elsewhere.[10] The analysis of 100 independent GA runs in Figure 2b depicts that the most informative ATS for the energy gap appear at a topological distance range of 1 to 42.

Details of the "best" topological SVM model to predict the $E_G$ of graphene appear in Table 1, while the rest of the GA-optimized machine learning predictors and further details of the optimum SVM model appear in the Supporting Information. It is worth noting that the "best" model in Table 1 has 17 topological distance inputs in a topological distance range from 1 to 87 and yields a cross-validation accuracy of ∼86%. This fact suggests that experimental characterization of graphene nanoflakes and probably other nanomaterials can be focused on a reduced set of topological distances across the topological structure; characterization of entire structures with atomic-level precision is not necessary.

**Table 1. Details of the Optimum SVM Model of the Energy Gap of Graphene Nanoflakes**

| topological distances | $R^2_{TFO}$ [a] | $R^2_{Test}$ [b] | $RMSE_{Test}$ (eV) [b] |
|---|---|---|---|
| 1, 3, 6, 7, 8, 11, 12, 16, 18, 20, 30, 45, 68, 75, 77, 84, 87 | 0.862 | 0.812 | 0.46 |

[a] $R^2_{TFO}$ is the squared Pearson's correlation coefficient of TFO cross-validation. [b] $R^2_{Test}$ and $RMSE_{Test}$ are the squared Pearson's correlation coefficient and root-mean-square error of the test set predictions, respectively.

The ability of the SVM model to predict the $E_G$ of the test set of 30% of the data set is illustrated in the scatter plot in Figure 3, and the squared correlation coefficient ($R^2_{Test}$) and root-mean-square error ($RMSE_{Test}$) of test set predictions appear in Table 1. Remarkably, the $R^2_{Test}$ value matched the cross-validation accuracy in Table 1 with a value higher than 0.81 and a low $RMSE_{Test}$ of ∼0.46 eV, which demonstrates that the machine learning model not only "learned" the topological pattern relevant for opening a gap but, more importantly, successfully generalized to new graphene structures.

The predictions of the highest values of $E_G$ in Figure 3 are less accurate, but the model correctly ranks the majority of their values. However, it is worth noticing that the topological approach exhibits test set accuracy that is ∼10 percent units higher than that of a partial least squares model of the interatomic distance distributions.[10] More importantly, with more than 80% accuracy, the topological predictions compare favorably to our recently reported machine learning model trained on a sophisticated combination of geometrical features, namely, aspect ratio, average carbon coordination number, edge conformation, and average bond distances and angles of graphene nanoflakes.[11]

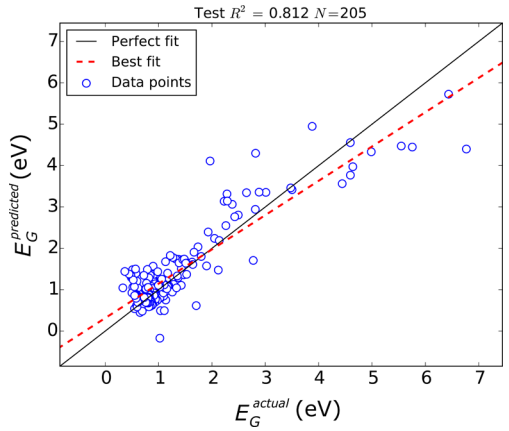



**Figure 3.** Scatter plot of the predictions for the energy gap of the graphene nanoflakes in the test set of 30% of the entire data set, where "actual" refers to the values calculated by DFTB simulations and "predicted" corresponds to the SVM predictions.

Considering that experimental comparison is not possible at the moment, our results for over 600 regular graphene nanoflakes with trigonal, hexagonal, and rectangular shapes and different degrees of passivation and edge conformation demonstrates that the size of the energy gap can be controlled to a large extent via the connectivity of the carbon atoms. This result suggests that different nanoflakes can be statistically discriminated into gap energy groups according to their topologies, which could guide engineering of nanoflake structures with specific topologies using, for example, scanning tunneling microscope lithography, which allows patterning of nanoribbons with well-defined widths and predetermined crystallographic orientations at nanometric precision.[19]

In summary, a machine learning model calibrated with topological information successfully predicted $E_G$ values for graphene nanoflakes with a squared correlation coefficient higher than 0.8 and an absolute error lower than 0.5 eV. The topological model can be useful to rapidly estimate the energy gap while providing some rationale for tuning the topology of graphene nanoflakes. To the best of our knowledge, this is the first machine learning prediction of the energy gap of graphene nanoflakes solely from molecular graph information. Furthermore, this approach could accelerate the development of graphene nanotechnologies by guiding the synthesis and/or lithographic preparation of graphene nanoflakes with moderated control over molecular connectivity and edge characteristics, and in general it can help experimentalists to identify rational strategies for the optimization of not only graphene but also other two-dimensional materials. In general, data-driven models can be integrated into in silico HT platforms for more efficient computational screening and analysis of the relationship between functional properties and structural features and imperfections of large nanomaterials libraries.



DOI: 10.1021/acscombsci.6b00094
ACS Comb. Sci. XXXX, XXX, XXX−XXX

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscombsci.6b00094.

Details of the DFTB calculations, machine learning implementation, calibration and testing, and outlier removal details (PDF)

ATS descriptors and $E_G$ values for the graphene data set (ZIP)

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: michael.fernandezllamosa@csiro.au. Phone: +61 3 9662 7151.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

Computational resources for this project were supplied by the Australian National Computing Infrastructure national facility under Grant q27.

## ■ REFERENCES


(1) Kosynkin, D. V.; Higginbotham, A. L.; Sinitskii, A.; Lomeda, J. R.; Dimiev, A.; Price, B. K.; Tour, J. M. Longitudinal unzipping of carbon nanotubes to form graphene nanoribbons. *Nature* **2009**, *458*, 872−876.

(2) Ritter, K. A.; Lyding, J. W. The influence of edge structure on the electronic properties of graphene quantum dots and nanoribbons. *Nat. Mater.* **2009**, *8*, 235−242.

(3) Park, S.; Ruoff, R. S. Chemical methods for the production of graphenes. *Nat. Nanotechnol.* **2009**, *4*, 217−224.

(4) Barnard, A. S.; Ōsawa, E. The impact of structural polydispersivity on the surface electrostatic potential of nanodiamond. *Nanoscale* **2014**, *6*, 1188−1194.

(5) Pedersen, T. G.; Flindt, C.; Pedersen, J.; Mortensen, N. A.; Jauho, A.-P.; Pedersen, K. Graphene antidot lattices: designed defects and spin qubits. *Phys. Rev. Lett.* **2008**, *100*, 136804.

(6) Son, Y. W.; Cohen, M. L.; Louie, S. G. Energy gaps in graphene nanoribbons. *Phys. Rev. Lett.* **2006**, *97*, 216803.

(7) Balog, R.; et al. Bandgap opening in graphene induced by patterned hydrogen adsorption. *Nat. Mater.* **2010**, *9*, 315−319.

(8) Berger, C.; Song, Z.; Li, X.; Wu, X.; Brown, N.; Naud, C.; Mayou, D.; Li, T.; Hass, J.; Marchenkov, A. N.; Conrad, E. H.; First, P. N.; de Heer, W. A. Electronic confinement and coherence in patterned epitaxial graphene. *Science* **2006**, *312*, 1191−1196.

(9) Chen, Z.-G.; Shi, Z.; Yang, W.; Lu, X.; Lai, Y.; Yan, H.; Wang, F.; Zhang, G.; Li, Z. Observation of an intrinsic bandgap and Landau level renormalization in graphene/boron-nitride heterostructures. *Nat. Commun.* **2014**, *5*, 4461.

(10) Fernandez, M.; Shi, H.; Barnard, A. S. Quantitative Structure-Property Relationship Modeling of Electronic Properties of Graphene Using Atomic Radial Distribution Function Scores. *J. Chem. Inf. Model.* **2015**, *55*, 2500−2506.

(11) Fernandez, M.; Shi, H.; Barnard, A. S. Geometrical features can predict electronic properties of graphene nanoflakes. *Carbon* **2016**, *103*, 142−150.

(12) Shi, H.; Barnard, a. S.; Snook, I. K. High throughput theory and simulation of nanomaterials: exploring the stability and electronic properties of nanographene. *J. Mater. Chem.* **2012**, *22*, 18119−18123.

(13) Barnard, A. Graphene Structure Set, v1. CSIRO Data Collection, 2014; DOI: 10.4225/08/541F61EC81EE3.

(14) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Model.* **1996**, *36*, 1205−1213.

(15) Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 834−852.

(16) Fernandez, M.; Ahmad, S.; Sarai, A. Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines. *J. Chem. Inf. Model.* **2010**, *50*, 1179−1188.

(17) Plachinda, P.; Rouvimov, S.; Solanki, R. Structure analysis of CVD graphene films based on HRTEM contrast simulations. *Proc. IEEE Conf. Nanotechnol.* **2011**, *2687*, 764−769.

(18) McNerny, D. Q.; Viswanath, B.; Copic, D.; Laye, F. R.; Prohoda, C.; Brieland-Shoultz, A. C.; Polsen, E. S.; Dee, N. T.; Veerasamy, V. S.; Hart, A. J. Direct fabrication of graphene on $SiO_2$ enabled by thin film stress engineering. *Sci. Rep.* **2014**, *4*, 5049.

(19) Tapasztó, L.; Dobrik, G.; Lambin, P.; Biró, L. P. Tailoring the atomic structure of graphene nanoribbons by scanning tunnelling microscope lithography. *Nat. Nanotechnol.* **2008**, *3*, 397−401.




DOI: 10.1021/acscombsci.6b00094
*ACS Comb. Sci.* XXXX, XXX, XXX−XXX