

ARTICLE OPEN

Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride

Yuan Dong¹, Chuhan Wu², Chi Zhang¹, Yingda Liu², Jianlin Cheng² and Jian Lin¹ 

It is well-known that the atomic-scale and nano-scale configuration of dopants can play a crucial role in determining the electronic properties of materials. However, predicting such effects is challenging due to the large range of atomic configurations that are possible. Here, we present a case study of how deep learning algorithms can enable bandgap prediction in hybridized boron–nitrogen graphene with arbitrary supercell configurations. A material descriptor that enables correlation of structure and bandgap was developed for convolutional neural networks. Bandgaps calculated by ab initio calculations, and corresponding structures, were used as training datasets. The trained networks were then used to predict bandgaps of systems with various configurations. For 4 × 4 and 5 × 5 supercells they accurately predict bandgaps, with a R^2 of >90% and root-mean-square error of ~0.1 eV. The transfer learning was performed by leveraging data generated from small supercells to improve the prediction accuracy for 6 × 6 supercells. This work will pave a route to future investigation of configurationally hybridized graphene and other 2D materials. Moreover, given the ubiquitous existence of configurations in materials, this work may stimulate interest in applying deep learning algorithms for the configurational design of materials across different length scales.

npj Computational Materials (2019)5:26; <https://doi.org/10.1038/s41524-019-0165-4>

INTRODUCTION

Introduction of defects at atomic or nanoscales has been a widely employed strategy in bulk materials, such as metals, ceramics, and semiconductors, for manipulation of their mechanical and physical properties. Especially when these defects are precisely controlled in configuration space, they significantly improve the mechanical properties,^{1–3} manipulate the magnetic properties,^{4–6} and alter the electronic properties.^{7–10} For instance, nanotwined grain boundaries are crucial to realize ultrahigh strength, superior fatigue resistance in metals,¹ and ultrahigh hardness and enhanced toughness in ceramics.^{2,3} Localization of nitrogen vacancy center at deterministic locations with nanoscale precision within diamond improves sensitivity and resolution of single electron spin imaging,^{4,5} and enhances charge-state switching rates,⁶ paving new ways to next-generation quantum devices. Precisely controlling dopant atoms in term of concentration and spatial arrangement within semiconductors is so critical to performance of fabricated electronic devices, especially as the dimension of the device keeps shrinking.⁷ Manipulation and detection of individual surface vacancies in a chlorine terminated Cu (100) surface realizes atomic scale memory devices.¹⁰

Based on the aforementioned examples in the bulk materials, it is naturally anticipated that the effect of the defects in two-dimensional (2D) materials, in term of the defect density and configurations, would be even more profound as all the atoms are confined within a basal plane with atomic thickness. Distinguished from allowance of different pathways of defect configurations in the three-dimensional (3D) bulk materials, this dimensionality restriction in the 2D materials largely reduces the accessibility and variability of the defects. This uniqueness would allow

configurational design of the defects in 2D, which starts to emerge as a new and promising research field. As the first well-known 2D material, graphene has been shown to exhibit configurational grain boundaries-dependent mechanical,^{11,12} thermal,¹³ and electrical properties.^{14,15} Doping graphene with heteroatoms, such as hydrogen, nitrogen, and boron, can further tailor the magnetic or electrical properties.¹⁶ Theoretical calculations suggest that these properties not only depend on types of dopants,^{16–18} and doping concentrations,¹⁹ but also are greatly determined by the dopants configurations within the graphene.^{20–22} Although the theoretical investigation already enables to research a much larger set of cases than the experiment does. The number of possible configurations for the dopants in graphene far exceeds the amount that can be practically computed due to extremely high computational cost. For instance, hybridizing boron–nitrogen (B–N) pairs into a graphene layer with a just 6 × 6 supercell system results in billions of possible configuration structures. Thus, it is entirely impractical to study all the possibilities to get the optimized properties even for such a small system. Another limitation of current mainstream of material design is that it heavily relies on intuition and knowledge of human who design, implement, and characterize materials through the trial and error processes.

Recent progress in data-driven machine learning (ML) starts to stimulate great interests in material fields. For instance, a series of material properties of stoichiometric inorganic crystalline materials were predicted by the ML.²³ In addition to their potentials in predicting properties of the materials, they start to show great power in assisting materials design and synthesis.^{24–27} It is anticipated that ML would assist to push the material revolution

¹Department of Mechanical & Aerospace Engineering, University of Missouri, Columbia, MO 65211, USA and ²Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, MO 65211, USA

Correspondence: Jianlin Cheng (chengji@missouri.edu) or Jian Lin (LinJian@missouri.edu)

These authors contributed equally: Yuan Dong, Chuhan Wu

Received: 8 October 2018 Accepted: 6 February 2019

Published online: 26 February 2019

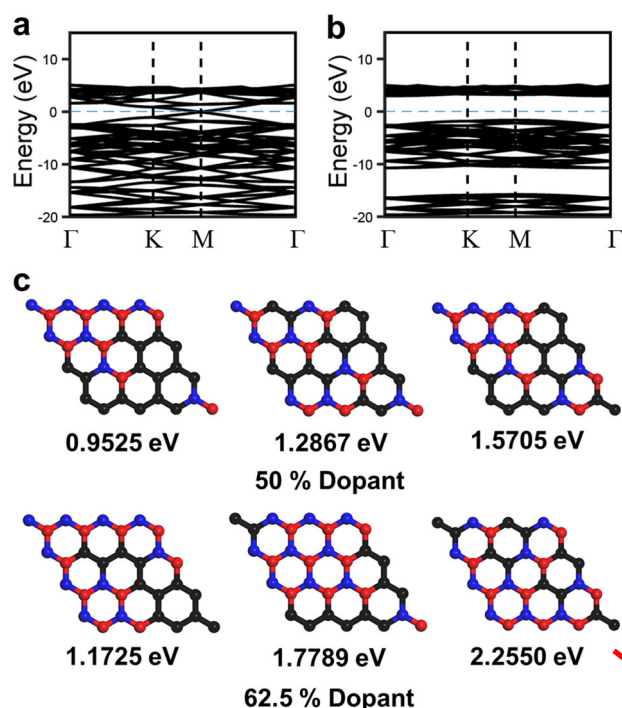


Fig. 1 **a, b** Band structures of pristine graphene **a** and pristine h-BN **b**. **c, d** Representative atomic configurations of 4×4 graphene supercell systems that have the same B-N dopant concentrations (50% and 62.5%) but with different configurational states and their corresponding bandgaps. C, N, B atoms are colored with black, blue, and red

to a paradigm of full autonomy in the next 5–10 years,^{28,29} especially as emerging of deep learning (DL) algorithms.^{30,31} For instance, a pioneer work of employing only a few layered convolutional neural networks (CNNs) enables to reproduce the phase transition of matters.³² Nevertheless, the application of the DL in the material fields is still in its infancy.³³ One of main barriers is that a compatible and sophisticated descriptive system that enables to correlate the predicted properties to structures is required for materials because DL algorithms are originally developed for imaging recognition.

Motivated by this challenge, we conceive to employ CNNs, including VGG16 convolutional network (VCN), residual convolutional network (RCN), and our newly-constructed concatenate convolutional network (CCN), for predicting electronic properties of hybridized graphene and h-BN with randomly configured supercells. As a benchmark comparison, the support vector machine (SVM),³⁴ which used to be the mainstream ML algorithm before the DL era, was also adopted (see details in Supplementary Note 1). We discovered that after trained with structural information and the bandgaps calculated from ab initio density function theory (DFT), these CNNs enabled to precisely predict the bandgaps of hybridized graphene and boron nitride (BN) pairs with any given configurations. One main reason for the high prediction accuracy arises from the developed material descriptor. Such a descriptive system enables to qualitatively and quantitatively capture the features of configurational states, where each atom in the structure affects its neighbor atoms so that these localized atomic clusters collectively determine bandgaps of the whole structure. Combined with well-tuned hyperparameters and well-designed descriptor, the CNNs result in high prediction accuracy. Considering that atom-scale precise structures of doped graphene by bottom-up chemical synthesis have been experimental realized,^{35–37} this work provides a cornerstone for future investigation of graphene and other 2D materials as well as their

associated properties. We believe that this work will bring up broader interests in applying the designed descriptive system and the CNN models for many materials related problems, which are not accessible to other ML algorithms.

RESULTS AND DISCUSSION

Dataset for bandgaps and structures of configurationally hybridized graphene

The generation of various configurations follows a pure geometric way. Because graphene and h-BN have similar honeycomb structures with very close bond lengths (1.42 Å and 1.44 Å for graphene and h-BN, respectively),³⁸ which is beneficial to structural stability if B–N pairs are hybridized in graphene. Moreover, graphene is a semimetal with a zero bandgap while h-BN is a wide-bandgap semiconductor. Thus, it can be naturally assumed that the graphene hybridized with the B–N pairs could have an intermediate bandgap. Finally, the B–N pairs can exactly make the charge neutral in the doped graphene. To test these hypotheses, we first implemented high throughput DFT calculations on configurationally hybridized graphene to generate datasets—correlation of supercell structures to bandgaps—for the ML. To achieve high-throughput calculations, we only apply non-hybrid function for the DFT calculations. Even though the calculated bandgaps are not as accurate as the previously reported values calculated by the hybrid functions, data consistence can be secured. Initial calculations show that the bandgaps of the pristine graphene with 4×4 , 5×5 , and 6×6 supercell systems are exactly 0 eV (Fig. 1a), and the bandgap of the pristine h-BN is 4.59 eV (Fig. 1b). More DFT calculations on examples of 4×4 systems that have the 50 at% B–N concentrations but with different configurational states show that they exhibit bandgaps ranging from 0.95 to 1.57 eV (Fig. 1c). As the BN concentration increases to 62.5 at%, the corresponding bandgaps also increase (Fig. 1d). The band structure of 6×6 supercell systems of graphene doped with 3 at% B–N concentrations (Fig. S1) illustrates the dopant induced bands/levels. These results validate that bandgaps of hybridized graphene depend on the configurations. Then more bandgaps of hybrids with arbitrary concentration and supercell configurations were calculated by the DFT. Some of them served as the training datasets for training the CNNs. Others served as test datasets to validate the accuracy of the prediction performed by the CNNs.

After the bandgap matrix is generated, structural information, such as chemical compositions and structures, needs to be well described in a form of numerical matrixes which will serve input data to train the CNNs. Defining descriptors of the materials for the ML is one of main challenges because the descriptors are more important to influence model accuracy than the ML algorithms do.^{23,39} Herein, we chose a simple and illustrative descriptor which is very suitable for the DL framework. We define the hybridized graphene structure by a 2D matrix described with only “0” and “1”. “0” corresponds to a C–C pair while “1” corresponds to a B–N pair (Fig. 2a). The size of matrix is 4×4 , 5×5 , and 6×6 , the same as the size of supercells. This natural representation like a 2D image significantly simplifies the learning process, securing the model accuracy. Note that we exclude the cases of switching configurations of B–N pairs, which will add more complexity to the investigated systems. The size of a matrix is determined by the size of a supercell in an investigated system. For instance, a 4×4 system was represented by a 4×4 matrix. In this work, 4×4 , 5×5 , and 6×6 systems were studied. To enlarge the training dataset, the equivalent structures were obtained by translating the particular structures along their lattice axis or inversion around their symmetry axis (Fig. S2). These structures are equivalent, thus having the same bandgaps. By this way, ~14,000, ~49,000, and ~7200 data examples for the 4×4 , 5×5 , and 6×6

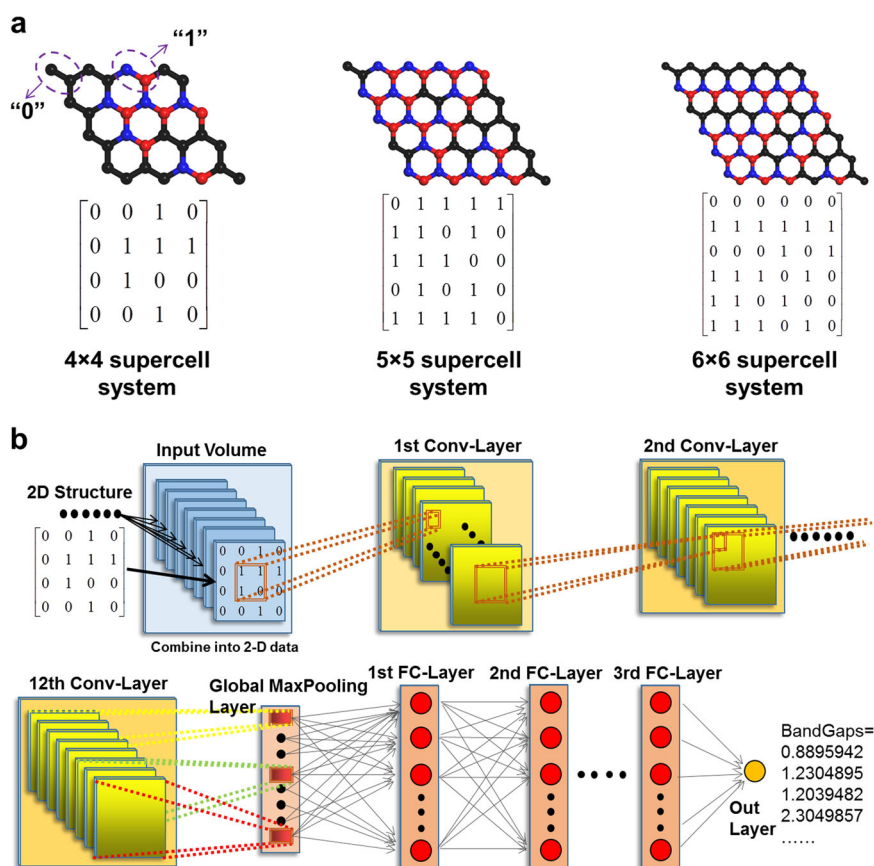


Fig. 2 **a** Descriptors for 2D doped graphene supercell systems (4 × 4, 5 × 5, and 6 × 6 systems). **b** A convolutional neural network, VCN, for the prediction of bandgaps of 2D doped graphene systems

6 supercell systems, respectively, were generated for purposes of training and validating CNNs. These datasets cover 21.36%, 0.15%, and 1×10^{-7} of all possible configurations for the 4 × 4, 5 × 5, and 6 × 6 supercell systems, respectively. These datasets were split into the training and test datasets, respectively. For example, when investigating 4 × 4 supercell systems, we randomly chose 13,000 and 1000 different data points from all the data samples as the training and validation datasets, respectively, to train and validate the CNN models. Correspondingly, for 5 × 5 supercell systems, the training and validation datasets are 48,000 and 1000, respectively. For 6 × 6 supercell systems, the training and validation datasets are 6200 and 1000, respectively.

Construction of CNNs

The general procedure of setting up the CNNs is illustrated as follows. The structural information of the hybridized graphene/BN sheets is represented by input matrices. The input matrices are transformed into multiple feature maps by filters in the first convolutional layer. The output of the convolutional layer is further transformed into high-level feature maps by the next convolutional layer. The maximum value of each feature map of the last convolutional layer is pooled together by the max pooling, which is used as the input of the next three fully-connected layers (FC). The single node in the output layer takes the output of the last FC layer as input to predict the bandgaps. The CNNs used in this work were constructed into three different structures. Firstly, we constructed a network which is similar to the traditional VCN designed for image processing,⁴⁰ as shown in Fig. 2b. This network has 12 convolutional (Conv) layers, one global-flatten layer, three FC layers, and an output layer. The neural layers in VCN are

explained in Supplementary Note 2. The detailed hyperparameters are given in Table S1.

Although the VCN is capable of learning from the data, our prediction results show that its performance is limited by the depth of neurons. The accuracy gets saturated and degraded rapidly. To tackle this problem, we further constructed the other two neural networks, RCN and CCN. The construction of RCN is explained in Supplementary Note 3. Its structure and hyperparameters are shown in Fig. S3 and Table S2–S4. The structure of RCN that we used is similar to that of ResNet50 network, but the Max Pooling layer was replaced with a Global Max pooling layer. The ResNet is a recent popular option in image classification.⁴¹ A characteristic of this network is that the convolution layers in VCN are replaced with residual blocks. This design can prevent the degeneration when the network goes deeper. The block in the CCN is our newly-constructed network which combines advantages of both GoogleNet⁴² and DenseNet.⁴³ The structure and hyperparameters of this network is explained in Supplementary Note 4 and illustrated in Fig. S4a and Table S5–S6. Similar to RCN, we also introduce the concatenation blocks into CCN as the building blocks (Fig. S4b and Table S6). Unlike RCN which “adds” feature maps in element-wise, CCN concatenates the layers from input and output data by simultaneously passing them through the activation layer. The concatenation of the network can also prevent it from degradation which is usually existing in the VCN. The concatenation can keep more extracted features in previous neuron layers, which would be beneficial to the transfer learning (TL). After the networks were built, they were trained by the generated datasets which correlate the structural information of the hybridized graphene with their corresponding bandgaps. After the CNNs were trained, 300 new datasets that illustrate 300

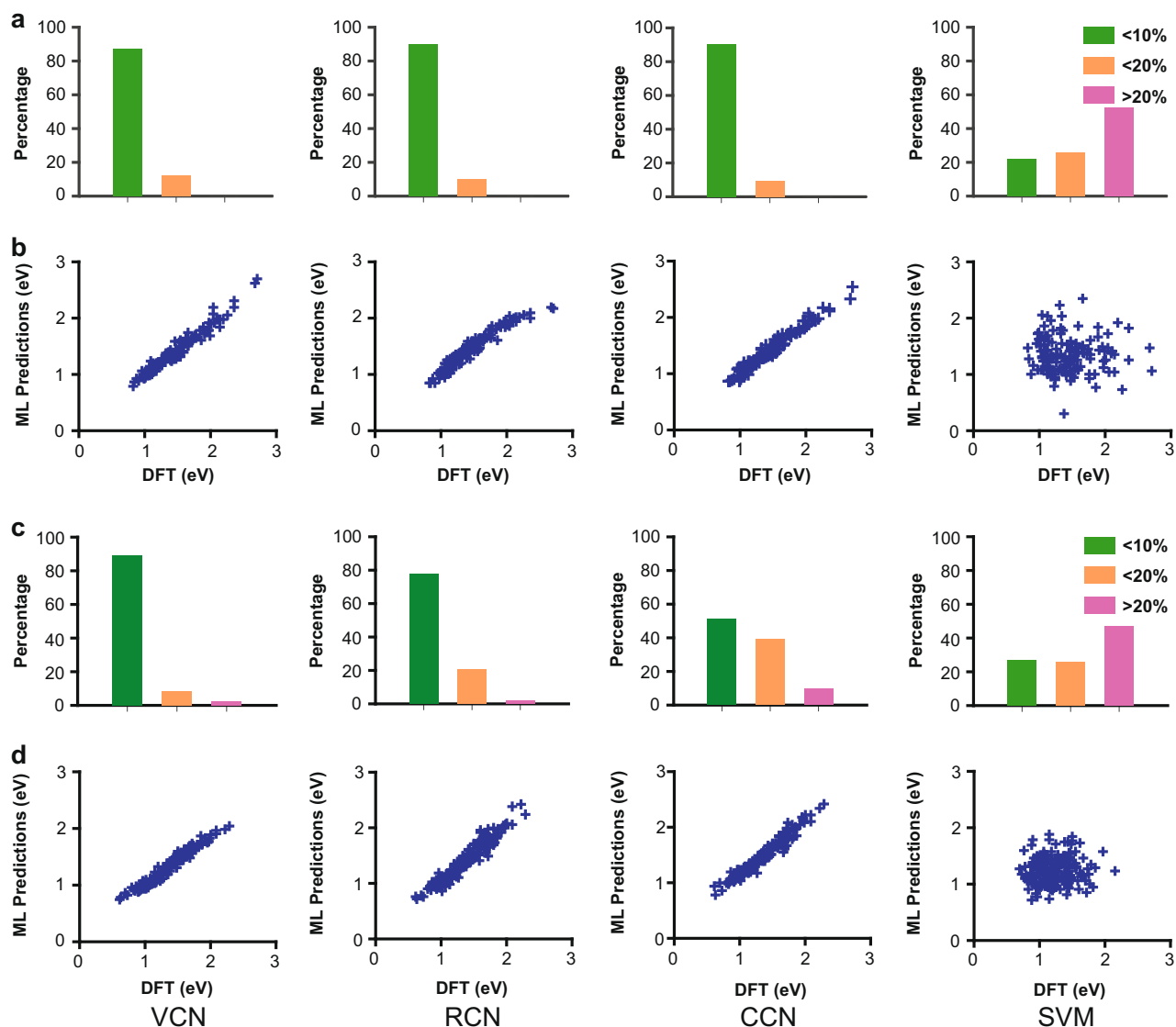


Fig. 3 Prediction performance of four ML algorithms. **a, c** Error levels of ML predicted bandgaps for 4×4 supercell systems **a** and 5×5 supercell systems **c**. The cases used for prediction are arranged by percentage of predicted cases showing prediction errors of <10%, <20%, and >20%. **b, d** ML predicted bandgaps vs. DFT calculated values for 4×4 supercell systems **b** and 5×5 supercell systems **d**

types of the hybridized graphene structures for each type of investigated supercell systems (4×4 , 5×5 , and 6×6) were fed into the trained CNNs to predict the bandgaps. Note that these used structures do not exist in ones used for training and validation. The predicted bandgaps were compared to the values calculated by the DFT on the same structures for evaluating the prediction accuracy of the DL algorithms.

Prediction of bandgaps by CNNs

As shown in Fig. 1, we can conclude that the bandgaps of the hybridized graphene are influenced by both the dopants and their configurations. In other words, each atom in the structure affects its neighbor atoms so that these localized atomic clusters collectively determine bandgaps of the whole structure. As convolutions in the CNNs can extract the features not only from the elements of the input data but also from their neighbors, it qualitatively and quantitatively captures the features of configurationally hybridized graphene, which will be proved as follows. The predicted bandgaps of the 4×4 and 5×5 supercell systems by different DL algorithms are compared with the results by the

DFT calculations (Fig. 3). Note that these data is obtained from the method of "learning from scratch" which suggests that the training and prediction are performed under the same graphene-h-BN hybrid systems.⁴⁴ For instance, the networks which are trained using the data from the 4×4 supercell systems are used to predict the systems with the same size but different configurations. The prediction accuracy is characterized by the relative error of the predicted bandgaps (E_{ML}) to the DFT calculated bandgaps (E_{DFT}), which is calculated as $|E_{ML} - E_{DFT}|/E_{DFT}$. As shown in Fig. 3a, all of three CNNs can predict the bandgaps of 4×4 supercell systems within 10% relative error for >90% cases. All of the predicted bandgaps for all cases have accuracy of >80%. In contrast, the prediction results from the SVM are deviated much more from the DFT benchmarks, showing >20% error for >50% cases. Figure 3b shows that the three CNNs exhibit strong direct linear correlation of ML predicted values and the DFT calculated values, while the SVM shows very weak correlation. The prediction accuracy of these CNNs for 5×5 supercell systems degrades a little (Fig. 3c). But the VCN network shows the prediction accuracy of >90% for >90% cases, which is the best among all three CNNs. The CCN has the lowest with prediction accuracy of >90% for only

Table 1. Statistics of predicted bandgaps by ML algorithms for 4×4 and 5×5 supercell systems

	MAE (eV)	MAE _F	RMSE (eV)	RMSE _F	R ²
4 × 4 systems					
VCN	0.08	5.34%	0.10	6.47%	0.9483
RCN	0.08	5.12%	0.11	6.44%	0.9426
CCN	0.07	4.59%	0.09	5.83%	0.9547
SVM	0.40	27.6%	0.33	34.7%	0.0029
5 × 5 systems					
VCN	0.07	5.72%	0.09	7.57%	0.9212
RCN	0.08	6.39%	0.10	8.33%	0.9124
CCN	0.15	13.3%	0.16	15.7%	0.9285
SVM	0.27	22.8%	0.51	29.5%	0.0032

~50% cases. That is possible due to lack of training data for 5×5 supercell systems considering their much larger configuration space than 4×4 supercell systems. The ML predicted bandgaps still have strong linear correlation with the DFT calculated ones (Fig. 3d). We believe that as the size of training datasets increases, the accuracy would be further improved. Similar to the prediction results shown for 4×4 supercell systems, when predicting bandgaps of 5×5 supercell systems the SVM shows poor performance (Fig. 3b, d).

In addition, other indicators of the prediction performance—the mean absolute error (MAE), root-mean-square error (RMSE) and explained variance (R²)—are provided in Table 1. Meanwhile, the fractional error, MAE_F and RMSE_F were also calculated, with their definition is shown in Supplementary Notes 5. For the 4×4 supercell systems, all three CNNs show very low RMSE of ~0.1 eV for the predicted bandgaps. The corresponding fractional errors, RMSE_F, for all three CNNs are ~6%. For the 5×5 supercell systems, the RMSE values slightly increase to 0.16 eV for the CCN, but it decreases to 0.09 eV and 0.10 eV when predicted by the VCN and RCN, further confirming the effectiveness of the VCN in predicting larger systems. The performance is also compared with different concentrations of BN pairs in the 5×5 supercell systems. Table S7 shows the statistics of prediction accuracy at three different levels of BN concentrations (<33 at%: “Low Concentration”; 33 at% ~ 66 at%: “Medium Concentration”; >66 at%: “High Concentration”). Compared with the results listed in Table 1, the accuracy doesn't change much, indicating the robustness of the CNN models in predicting the bandgaps of the structures with different BN concentrations.

The ultralow RMSE and RMSE_F values show that these DL algorithms are more effective in predicting bandgaps of our system than other material systems, such as double perovskites³⁹ which show a RMSE of 0.36 eV or inorganic crystals which show a RMSE of 0.51 eV.²³ This advantage is even more compelling if considering that: (i) our performance is rigorously evaluated with the newly generated systems that don't have any translational or symmetry equivalence with the training ones; (ii) the error maintains a low level when the relative size of the training data is significantly reduced for the 5×5 supercell systems. In contrast, the prediction accuracy from the SVM algorithm is lower, showing higher RMSEs of 0.33 and 0.51 eV for the 4×4 and 5×5 supercell systems, respectively. The R² is an indicator of correlation between the prediction and real values, which is considered as one of the most important metrics for evaluating the accuracy of the prediction models. Table 1 illustrates that the predicted bandgaps by all three CNNs have ~95% and >90% relevance to the values calculated by the DFT for the 4×4 and 5×5 supercell systems, respectively. Among them, the RCN shows the best prediction results. In contrast, the SVM has a near zero R², indicating almost

no relevance between the two. In summary, these results show that the CNNs are superior to non-CNN ML algorithms in predicting bandgaps of the configurationally hybridized graphene. People generally agree that the CNN methods perform better than non-CNN ones in terms of feature extraction for problems involving spatial structures. This advantage could be attributed to the convolution processing of data in CNNs.^{45–47} The flattening process of SVM method could lose important spatial features which are deterministic to bandgaps. Therefore, our results presented in this work provide another example for the effectiveness of CNNs dealing with spatial problems in material science.

Transfer learning: training and prediction

As suggested from the prediction results shown in the 4×4 and 5×5 supercell systems, the prediction accuracy is decreased as the relative size of training data shrinks. Obtaining sufficient training data, such as bandgaps calculated by the DFT, can lead to unusually high cost especially as the system scales up. Such a problem imposes a major challenge in the application of ML to the materials science. To overcome this challenge, an emerging TL has been proposed.⁴⁸ To conceptually validate the effectiveness of the TL in predicting bandgaps of larger systems we leveraged relatively larger datasets generated from the 4×4 and 5×5 supercell systems to improve models trained on more limited datasets generated in the 6×6 supercell systems. To do that, we built TL frameworks based on the CCN, RCN, and VCN. These networks were trained with the datasets previously used for training the 4×4 and 5×5 supercell systems together with 7200 new data points generated from the 6×6 supercell systems. The TL procedures for all CNNs are similar to the ones shown in refs.^{49,50} where all CNNs layers except the last FL are set at a learning rate 10% of the original learning rate. The last layer is re-normalized and trained with the new dataset. Its learning rate is set to 1% of the original CNN networks.

Figure 4a shows prediction errors of bandgaps for the 6×6 supercell systems by all three DL algorithms without the TL. Compared with the 4×4 and 5×5 systems they are significantly increased in all three categories of the prediction errors. The VCN shows >90% prediction accuracy for only ~40% cases. The RCN performs the best among all three ones with near 55% cases reaching >90% accuracy. The CCN shows the lowest percentage (i.e., ~45%) of cases within 10% prediction error and has 15% of the cases with >20% error. Figure 4b shows that the correlation between the predicted bandgaps and the DFT calculated ones is weaker for all three CNNs compared with the smaller systems, due to the much smaller percentage of sample size (only 1×10^{-7} of all possibilities). The prediction accuracy can be notably boosted after they are combined the TL (Fig. 4c). The CCN with the TL performs the best, with >60% cases achieving >90% accuracy. The percentage of cases with >20% error reduces to <10% from 20%. It is a significant improvement considering that increasing prediction accuracy becomes more and more difficult after a certain point. With the TL, the RCN and VCN also shows improved prediction accuracy. The percentage of the cases that show >20% error decreases from 20 to 8% when using the RCN, while VCN decrease percentage of cases from 35 to 12%. The correlation between the predicted bandgaps and the DFT calculated ones becomes much stronger after using the TL methods (Fig. 4d). The statistics of the predicted bandgaps by the different DL algorithms with and without the TL for the 6×6 supercell systems is shown in Table 2. Overall, the TL boosts the prediction accuracy of all three DL algorithms in terms of reducing the MAE and RMSE. For instance, it helps to reduce the MAE of CCN from 0.13 to 0.09 eV and RMSE from 0.16 to 0.12 eV. Their RMSE_F values are also reduced. This accuracy is comparable to the prediction accuracy show in the 5×5 supercell systems with the same DL algorithms.

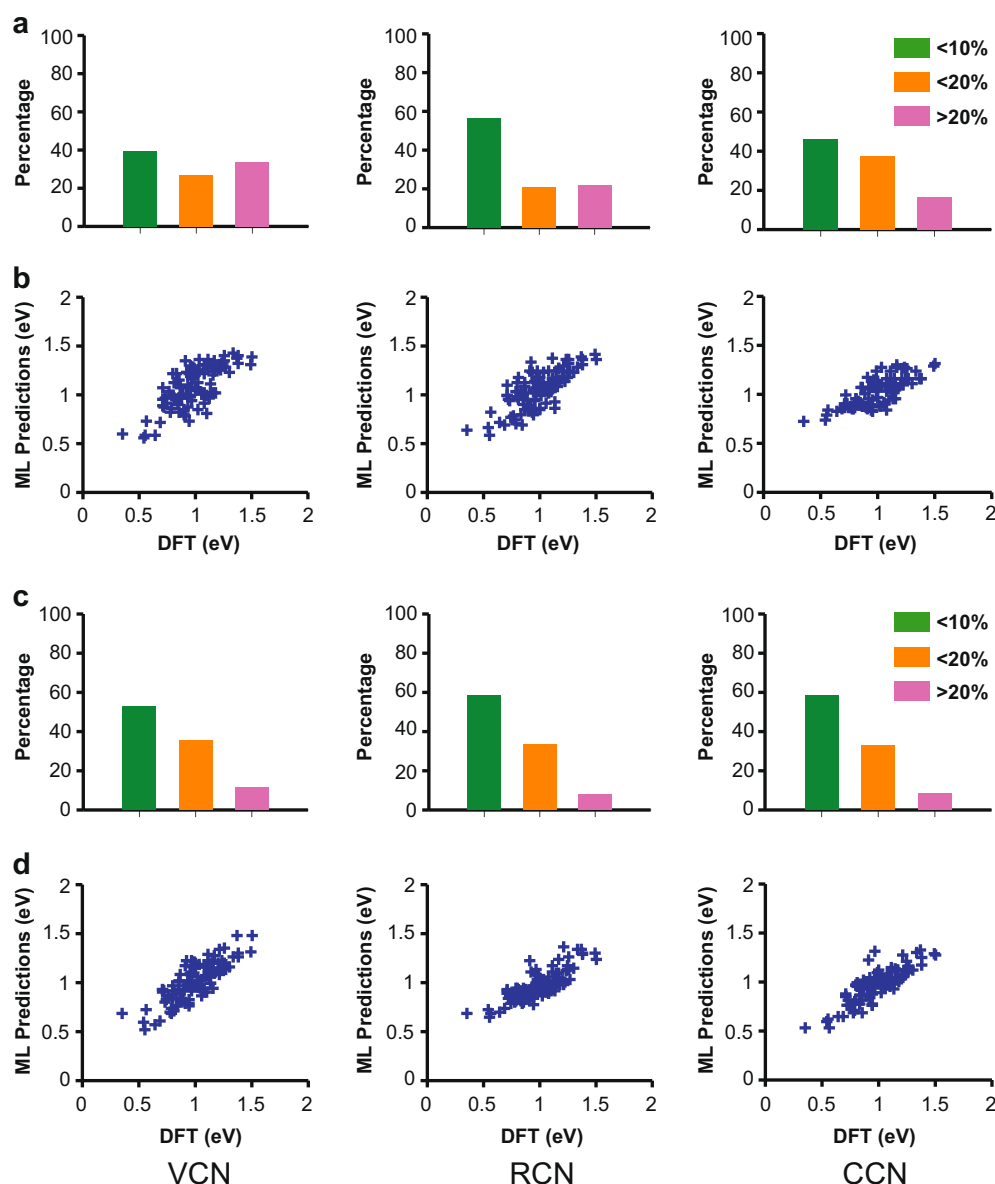


Fig. 4 Prediction performance of three DL algorithms **a, b** before and **c, d** after transfer learning for 6×6 supercell systems. **a, c** Error levels of ML predicted bandgaps before TL **a** and after TL **c**. The cases used for prediction are arranged by percentage of predicted cases showing prediction errors of <10%, <20%, and >20%. **b, d** ML predicted bandgaps vs. DFT calculated values before TL **b** and after TL **d**

Table 2. Statistics of predicted bandgaps by different DL algorithms without and with transfer learning for 6×6 supercell systems

	MAE (eV)	MAE _F	RMSE (eV)	RMSE _F	R^2
<i>Without transfer learning</i>					
VCN	0.15	15.7%	0.18	20.6%	0.5390
RCN	0.12	13.6%	0.15	19.4%	0.5243
CCN	0.13	14.5%	0.16	19.8%	0.5771
<i>With transfer learning</i>					
VCN	0.11	11.3%	0.14	15.5%	0.5958
RCN	0.10	10.2%	0.12	15.0%	0.6525
CCN	0.09	9.60%	0.12	12.6%	0.6642

They are the lowest among the values predicted by all three CNNs used in this work. This demonstration of applying the TL to predict bandgaps of the configurationally hybridized graphene of larger size would pave new route to mitigating barriers for the ML in solving challenges of data scarcity faced in the material fields.

In summary, we develop the DL models to predict the bandgaps of hybrids of graphene and h-BN with arbitrary supercell configurations. Three CNNs yield high prediction accuracy (>90%) for 4×4 and 5×5 supercell systems. The CNNs show superior performance to the non-CNN ML algorithms. The TL, leveraging the pre-trained network on small systems, boosts the prediction accuracy of three CNNs when predicting the bandgaps of large systems. The resulting MSE and RMSE of the predicted bandgaps of 6×6 systems by the CCN can be close to those of the 5×5 systems, but with a much smaller sampling ratio. Through this scientifically significant example, we have successfully illustrated the potential of artificial intelligence in studying 2D materials with various configurations, which would pave a new

route to rational design of materials. Given the recent progress in atom-scale precise structures of 2D materials realized by bottom-up chemical synthesis, we anticipate that our methods could form a computational platform that enables to pre-screen candidates for experimental realization. Moreover, due to the extremely high computation cost from the DFT calculation for larger systems,⁵¹ this platform could be useful to extrapolate the results obtained from smaller systems to larger systems with assistance of the TL algorithm. Upon the success of CNNs in present material systems, we envision that they could be extendable to 2D material systems with different number of layers considering that the CNN models are also effective in RGB images classification by preprocessing the images into multiple 2D numerical matrices.⁵² In this case, the images can be encoded to 2D numerical matrices and then fed into the CNN models. Nevertheless, this hypothesis requires careful investigation due to layer-to-layer interaction resulted physical and chemical complexity. Nevertheless, we believe that current work will inspire researchers in 2D materials for further exploring this promising area.

METHODS

DFT calculation

The ab initio DFT calculations were performed by **QUANTUM ESPRESSO package**.⁵³ We have successfully employed DFT calculations to investigate the nitrogen doping in graphene and to predict a class of novel two-dimensional carbon nitrides.^{54,55} The ultra-soft projector-augmented wave (PAW) pseudopotential⁵⁶ was used to describe the interaction between the valence electrons and the ions. The Perdew-Burke-Ernzerhof (PBE) function was applied for the exchange-correlation energy.⁵⁷ The cutoff plane wave energy was set to 400 eV. The Monkhorst-Pack scheme⁵⁸ was applied to sample the Brillouin zone with a mesh grid from $12 \times 12 \times 1$ in the k-point for all the systems. The graphene sheets were modeled as 2D matrixes. The matrix with all zero represents the **intrinsic graphene**, and the matrix with all one represents the **intrinsic h-BN**. The hybridized graphene and h-BN sheets were modeled by matrixes with 0 and 1 elements. The sizes of computed 2D sheets were 4×4 , 5×5 , and 6×6 supercells. The output training datasets were the bandgaps of randomly generated 2D graphene and BN hybrids with random concentration of BN pairs varying from 0 to 100 at% (Fig. S5). Specifically, the configurations were generated by random sampling. For example, for the 4×4 systems, we generated a set of random decimal numbers between 0 and 2¹⁶. Then these decimal numbers were converted into binary numbers that were represented with 16 digits. These digits were further converted to 4 by 4 matrixes to represent samples showing various configurations.

DATA AVAILABILITY

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

ACKNOWLEDGEMENTS

J.L. acknowledges financial support from University of Missouri-Columbia start-up fund, NASA Missouri Space Consortium (Project: 00049784), United States Department of Agriculture (Award number: 2018-67017-27880). This material is based upon work partially supported by the Department of Energy National Energy Technology Laboratory under Award Number DE-FE0031645. J.C. acknowledges National Science Foundation (Award numbers: DBI1759934 and IIS1763246). The computations were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC), supported in part by NSF (award number: 1429294).

AUTHOR CONTRIBUTIONS

J.L. conceived the project. Y.D. designed material descriptive system, performed the DFT calculations, and did the data analysis. C.Z. assisted Y.D. in DFT calculation and data analysis. J.C. designed the deep learning methods. C.W. built the CNNs and performed the training and prediction assisted by Y. Liu. J.L. and J.C. supervised the project. Y.D. and C.W. contributed equally to this work. All authors contributed to the discussions and writing of the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the npj Computational Materials website (<https://doi.org/10.1038/s41524-019-0165-4>).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Lu, K., Lu, L. & Suresh, S. Strengthening materials by engineering coherent internal boundaries at the nanoscale. *Science* **324**, 349–352 (2009).
- Tian, Y. J. et al. Ultrahard nanotwinned cubic boron nitride. *Nature* **493**, 385–388 (2013).
- Huang, Q. et al. Nanotwinned diamond with unprecedented hardness and stability. *Nature* **510**, 250–253 (2014).
- Mamin, H. J. et al. Nanoscale nuclear magnetic resonance with a nitrogen-vacancy spin sensor. *Science* **339**, 557–560 (2013).
- Grinolds, M. S. et al. Subnanometre resolution in three-dimensional magnetic resonance imaging of individual dark spins. *Nat. Nanotechnol.* **9**, 279–284 (2014).
- Doi, Y. et al. Deterministic electrical charge-state initialization of single nitrogen-vacancy center in diamond. *Phys. Rev. X* **4**, 011057 (2014).
- Shinada, T., Okamoto, S., Kobayashi, T. & Ohdomari, I. Enhancing semiconductor device performance using ordered dopant arrays. *Nature* **437**, 1128–1131 (2005).
- Kitchen, D., Richardella, A., Tang, J. M., Flatte, M. E. & Yazdani, A. Atom-by-atom substitution of Mn in GaAs and visualization of their hole-mediated interactions. *Nature* **442**, 436–439 (2006).
- Koenraad, P. M. & Flatte, M. E. Single dopants in semiconductors. *Nat. Mater.* **10**, 91–100 (2011).
- Kalff, F. E. et al. A kilobyte rewritable atomic memory. *Nat. Nanotechnol.* **11**, 926–929 (2016).
- Grantab, R., Shenoy, V. B. & Ruoff, R. S. Anomalous strength characteristics of tilt grain boundaries in graphene. *Science* **330**, 946–948 (2010).
- Shekhawat, A. & Ritchie, R. O. Toughness and strength of nanocrystalline graphene. *Nat. Commun.* **7**, 10546 (2016).
- Ma, T. et al. Tailoring the thermal and electrical transport properties of graphene films by grain size engineering. *Nat. Commun.* **8**, 14486 (2017).
- Tsen, A. W. et al. Tailoring electrical transport across grain boundaries in polycrystalline graphene. *Science* **336**, 1143–1146 (2012).
- Fei, Z. et al. Electronic and plasmonic phenomena at graphene grain boundaries. *Nat. Nanotechnol.* **8**, 821–825 (2013).
- Cruz-Silva, E., Barnett, Z. M., Sumpter, B. G. & Meunier, V. Structural, magnetic, and transport properties of substitutionally doped graphene nanoribbons from first principles. *Phys. Rev. B* **83**, 155445 (2011).
- Martins, T. B., Miwa, R. H., da Silva, A. J. R. & Fazzio, A. Electronic and transport properties of boron-doped graphene nanoribbons. *Phys. Rev. Lett.* **98**, 196803 (2007).
- Kim, S. S., Kim, H. S., Kim, H. S. & Kim, Y. H. Conductance recovery and spin polarization in boron and nitrogen co-doped graphene nanoribbons. *Carbon N. Y.* **81**, 339–346 (2015).
- Lherbier, A., Blase, X., Niquet, Y. M., Triozon, F. & Roche, S. Charge transport in chemically doped 2D graphene. *Phys. Rev. Lett.* **101**, 036808 (2008).
- Yu, S. S., Zheng, W. T., Wen, Q. B. & Jiang, Q. First principle calculations of the electronic properties of nitrogen-doped carbon nanoribbons with zigzag edges. *Carbon N. Y.* **46**, 537–543 (2008).
- Biel, B., Blase, X., Triozon, F. & Roche, S. Anomalous doping effects on charge transport in graphene nanoribbons. *Phys. Rev. Lett.* **102**, 096803 (2009).
- Zheng, X. H., Rungger, I., Zeng, Z. & Sanvito, S. Effects induced by single and multiple dopants on the transport properties in zigzag-edged graphene nanoribbons. *Phys. Rev. B* **80**, 235426 (2009).
- Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
- Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Xue, D. Z. et al. Accelerated search for BaTiO₃-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proc. Natl Acad. Sci. USA* **113**, 13301–13306 (2016).
- Xue, D. Z. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
- Nikolaev, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 16031 (2016).
- Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation, *Nature Reviews. Materials* **3**, 5–20 (2018).

29. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
30. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
31. Gibney, E. Google AI algorithm masters ancient game of Go. *Nature* **529**, 445–446 (2016).
32. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
33. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
34. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
35. Kawai, S. et al. Atomically controlled substitutional boron-doping of graphene nanoribbons. *Nat. Commun.* **6**, 8098 (2015).
36. Nguyen, G. D. et al. Atomically precise graphene nanoribbon heterojunctions from a single molecular precursor. *Nat. Nanotechnol.* **12**, 1077–1082 (2017).
37. Kawai, S. et al. Multiple heteroatom substitution to graphene nanoribbon. *Sci. Adv.* **4**, 7181 (2018).
38. Jin, C., Lin, F., Suenaga, K. & Iijima, S. Fabrication of a freestanding boron nitride single layer and its defect assignments. *Phys. Rev. Lett.* **102**, 195505 (2009).
39. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
40. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition, *arXiv*, arXiv: 1409.1556 [cs.CV], (2014).
41. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision.* **115**, 211–252 (2015).
42. Szegedy, C. et al. Going deeper with convolutions. *Proc. IEEE Conf. Comput. Vision. Pattern Recognit.* **2015**, 1–9 (2015).
43. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *CVPR* **2017**, 3 (2017).
44. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
45. Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. CNN features off-the-shelf: an astounding baseline for recognition. *Proc. IEEE Conf. Comput. Vision. Pattern Recognit. Workshops* **2014**, 806–813 (2014).
46. Liu, Z., Hu, J., Weng, L. & Yang, Y. Rotated region based CNN for ship detection. *Image Processing (ICIP). 2017 IEEE Int. Conf.* **2017**, 900–904 (2017).
47. Yin, W., Kann, K., Yu, M. & Schütze, H. Comparative study of cnn and rnn for natural language processing. *arXiv Prepr. arXiv* **1702**, 01923 (2017).
48. Hutchinson, M. L. et al. Overcoming data scarcity with transfer learning. *arXiv*, arXiv **1711**, 05099 (2017).
49. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE. Trans. Pattern Anal. Mach. Intell.* **38**, 142–158 (2016).
50. Hoo-Chang, S. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285 (2016).
51. Ci, L. et al. Atomic layers of hybridized boron nitride and graphene domains. *Nat. Mater.* **9**, 430 (2010).
52. CireşAn, D., Meier, U., Masci, J. & Schmidhuber, J. Multi-column deep neural network for traffic sign classification. *Neural Netw.* **32**, 333–338 (2012).
53. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **21**, 395502 (2009).
54. Dong, Y., Gahl, M. T., Zhang, C. & Lin, J. Computational study of precision nitrogen doping on graphene nanoribbon edges. *Nanotechnology* **28**, 505602 (2017).
55. Dong, Y., Zhang, C., Meng, M., Melinda, G. & Lin, J. Novel two-dimensional diamond like carbon nitrides with extraordinary elasticity and thermal conductivity. *Carbon* **138**, 319–324 (2018).
56. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
57. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
58. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188 (1976).



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019