Accurate energy prediction of large-scale defective two-dimensional materials via deep learning

View Online   Export Citation   CrossMark

Yuxing Ma, Shuaihua Lu, Yehui Zhang, Tingbo Zhang, Qionghua Zhou,[a] and Jinlan Wang[a] (iD)

AFFILIATIONS

School of Physics, Southeast University, Nanjing 211189, China

[a]Authors to whom correspondence should be addressed: qh.zhou@seu.edu.cn and jlwang@seu.edu.cn

ABSTRACT

Crystal defects play a vital role in physical and chemical properties of two-dimensional (2D) materials. The computational cost for a real defective system with more than thousands of atoms is considerably high. Here, we propose a framework to accurately predict the formation energy of defective 2D materials, graphene and molybdenum disulfide (MoS_2), at a large scale, by combining deep learning techniques and density functional calculations. To improve the training performance of deep learning models, a multi-layer structure descriptor using chemical bond parameters is proposed. For the defective graphene (MoS_2) over 300 nm^2 (650 nm^2), the calculated mean absolute error for the formation energy is less than 47 meV (53 meV) per 1000 atoms. This study provides a practical solution for the accurate and rapid description of large-scale defective 2D materials.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0091994

During the growth and preparation of two-dimensional (2D) materials, a certain number of defects and disorder will inevitably occur. Defects can change the topological structure of 2D materials, greatly affecting the electronic, chemical, mechanical, and thermal properties of materials.[1–5] Experimentally, characterization techniques like scanning transmission electron microscopy (STEM) and Raman spectroscopy can provide general information of 2D materials with defects, but describing and understanding at the atomic-scale is hard to reach by experimental tools alone. Calculations based on first-principles methods like density functional theory (DFT) are commonly used to serve this purpose. Nevertheless, considering the computational cost, only small systems up to a few hundred atoms can be handled in principle. When dealing with 2D materials with multiple defects or low defect concentration (which will need a much larger simulation cell and easily over 1000 atoms), the cost of *ab initio* calculations would be unbearable if not unreachable. Clearly, to understand the large-scale defective 2D system better, a new approach is needed, which should be cheap, fast, and accurate at the DFT level.

In recent years, machine learning (ML) has been developed rapidly to accelerate material design. In particular, deep learning (DL) based on an artificial neural network is deemed to be more accurate for big data, since the multiple hidden layers have higher learning capabilities and can handle large data sets and complex relationships better. DL has been successfully applied in material property prediction,[6–13] searching for new materials,[14–17] structural stability analysis,[18,19] and so on. For example, the DL-based method Crystal Graph Convolutional Neural Network (CGCNN) learns material properties from the connections of atoms in the crystal and has achieved good results in the prediction of formation energy, bandgap, and total magnetization.[20] Wang *et al.* predicted the Heyd–Scuseria–Ernzerhof (HSE06) bandgaps of 1503 binary metal oxides, nitrides, and sulfides by the DL model, which is at least 10^4 times faster than DFT calculations.[15] Hundi *et al.* developed a descriptor based on the voxelization method, which can help DL models to achieve good prediction accuracy at lower computational cost.[21] For large-scale 2D defective systems, which have massive number of atoms, multiple defect types, and complex defect distribution, DL would be an appropriate choice. However, there are still some challenges. Data are always the first and biggest one. Although there are large databases of 2D materials, such as Material Project,[22] OQMD,[23] and C2DB,[24] their data are usually limited to perfect materials, and data sets of defective systems are still rare. The second challenge is proper descriptor. Descriptors that can describe the local atomic environment are needed, specifically for the defective systems. A proper descriptor should be able to distinguish different types of defects and different defect concentrations and distributions. For most descriptors, with the increase in the number of atoms, the dimension of the descriptor will also increase, which may lead to a dimensional disaster.[25] Therefore, a method that can reduce the dimension of the descriptor without losing information is required.

In this work, we develop a convolutional neural network (CNN) to predict the formation energy of large-scale defective systems, since the formation energy of defects is a key metric of its formability and stability. To accurately describe the structure of different defect types and the distribution of defects, a multi-layer structure descriptor is proposed. By encoding the chemical bond parameters in 2D materials, the structure can be completely described. Defective graphene is chosen as an example for our CNN workflow. Model accuracy at different defect concentrations, sensitivity to defects distances, and generalization ability to unknown defects are comprehensively tested. The impact of defects types, concentrations, and distributions on the total formation energy of the whole system is further studied. Our CNN model training with the multi-layer descriptor can accurately predict the formation energy of the graphene system over $300\,nm^2$ with the mean absolute error (MAE) less than $47\,meV$ per 1000 atoms. This method is also extended to more complicated two-dimensional defective systems like molybdenum disulfide (MoS_2) that the MAE of the predicted formation energy of MoS_2 system over $650\,nm^2$ is less than $53\,meV$ per 1000 atoms.

Considering the complexity, we take the relatively simple system, graphene as an example. A unique feature of graphene lattice defects is that it can be reconstructed through non-hexagonal shapes. For example, Stone–Wales (SW) defects are formed by rotating one of the C–C bonds in perfect graphene by 90, and four hexagons are transformed into two pentagons and two heptagons [SW (55–77)],[26] as shown in Fig. 1(a). Thus, whether it is a single-vacancy defect or a multi-vacancy defect, in essence, the bond position is changed on the basis of

the perfect graphene lattice, or the bond is stretched and rotated, which causes the distortion of the crystal lattice. This provides a theoretical basis for the establishment of our descriptors.

We propose a three-layer descriptor, in which the chemical bonds between atoms are selected as the description unit, and the defective graphene system is described by utilizing the voxelization method to construct a three-dimensional matrix descriptor, as shown in Fig. 1(b). The chemical construction parameter matrices consist of bond position matrix (BPM), bond length matrix (BLM), and bond angle matrix (BAM). The nature of voxelization approach determined that the bond-centered descriptor is more sensitive to the break of bonds and reconstruction around defects than the atom-centered descriptor. For example, the multi-layer descriptor contains bond position information, which can identify single-vacancy and multi-vacancy defects, and bond length and bond angle information, which can identify SW defects. The advantage of this multi-layer descriptor is that it can describe the situations where the surrounding atoms deviate from their original position due to defects. From the visualization of BLM and BAM, we can see that the distortion in the length and angle around defects is significant, as shown in Fig. 1(c). In places far away from defects, there is negligible distortion. More details about the descriptor can be found in the supplementary material.

To obtain the energy of a large-scale graphene defect system at low cost and quickly, we designed a multi-step workflow, as shown in Fig. 2. First, the formation energies of different defects in the small unit cell are obtained by DFT calculations and the influences of the interaction between defects on energy are also considered, as shown in
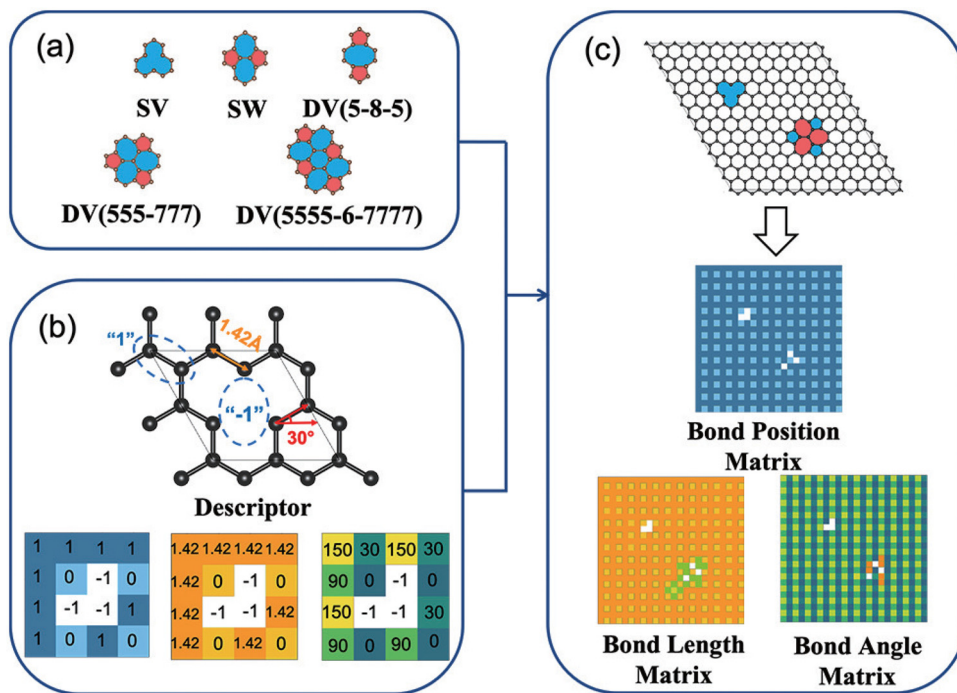

FIG. 1. Schematic diagram of the multi-layer structure descriptor. (a) All types of defects in the data set. It includes single vacancy (SV), Stone–Wales (SW), and double vacancy (DV), which are composed of multiple different polygons. (b) Multi-layer matrix descriptor coding rules. (c) Descriptor visualization diagram after structural encoding. The green element block in bond length matrix (BLM) and the orange element block in bond angle matrix (BAM) indicate that the value of the bond parameter has changed due to lattice distortion and does not represent the specific value.
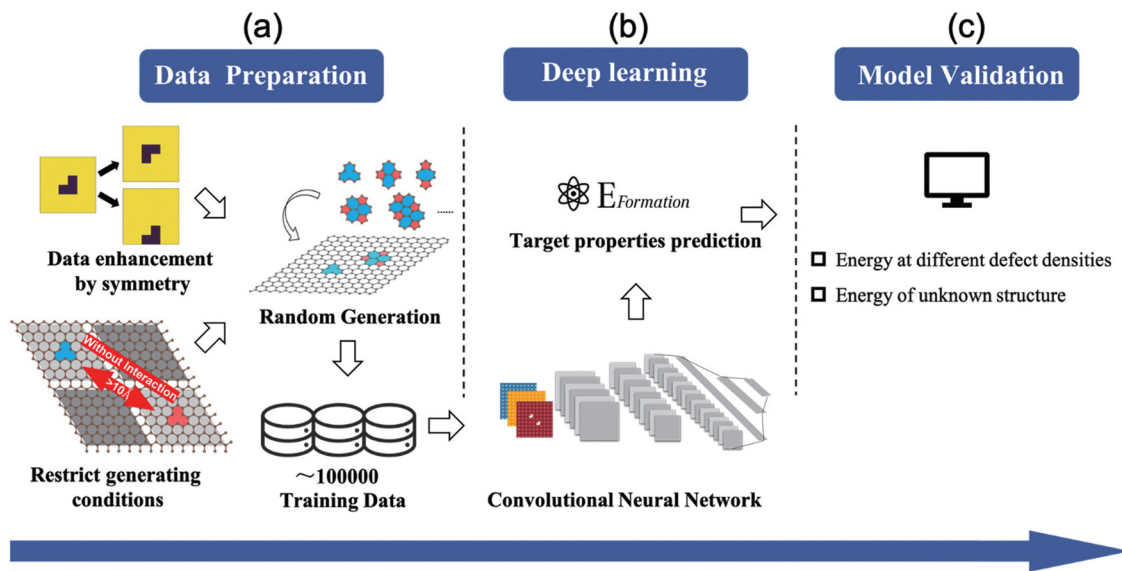
FIG. 2. Work flow for predicting formation energy of large-scale defective graphene. (a) Data preparation. Generate large amounts of data through random generation under restricted conditions. The database is built by translating and rotating the original data. (b) Deep learning process. We use the convolutional neural network model. (c) Model verification. Perform index evaluation on the model to verify the reliability and stability of the model. Use the model to predict defects and unknown structures at different concentrations, and then perform DFT verification.

Fig. 2(a). We compare the formation energy of individual defects with previous reports (see details in Table S2). Moreover, DFT results show that the interaction between different defects has a non-negligible effect on energy. As the distance of interaction increases, some of the energy increases monotonously and some decreases monotonously, more specific details are shown in Table S3. For the distance between defects greater than 15 Å, the interaction is very weak, and the impact on energy is negligible. Therefore, when we construct a large system, we can obtain the overall energy by adding a series of small cell energies. We only need to ensure that the distance between defects in different cells is greater than 15 Å, and the defects with a distance less than 15 Å are placed in the same cell. In this way, we can obtain large amounts of data for CNN training at low cost. We put the above-mentioned different types of structures into the data pool, and under the condition that the distance between the defect combination and the surrounding graphene sheets is larger than 15 Å, we randomly select them and then splice them to ensure the randomness and diversity of the structure. In this way, we produce the first generation of raw data.

Since the structures that we consider have periodic boundary conditions, we perform translation and rotation operations on the structure description matrix along a certain lattice axis to obtain the equivalent structure of the first-generation structure. We use this method to generate second-generation data to increase the training data. In order to make the data set more complete and increase the robustness of our model, based on the above random generation method, we generate graphene data with different defect concentrations that conform to the Gaussian distribution (Fig. S1).

Due to the huge data dimensions brought by the large system, the convolutional neural network (CNN) is chosen to predict the formation energy. Through feature extraction operations, such as

convolution and pooling, the dimensionality is reduced, as shown in Fig. 2(b). More details about the topology of CNN are given in Fig S2. As shown in Fig. S5, the process of convolution can be seen through the visual feature map under the convolution and pooling operation. The network extracts information, such as the distribution status of each individual defect and the characteristics of the local environment around the defect, and learns the mapping relationship between the defect position and the formation energy. Finally, the reliability of our model is verified through various evaluation indicators of the model, as shown in Fig. 2(c).

As shown in Fig. 3(a), the coefficient of determination (R^2) of the CNN model is 0.998 and the mean absolute error (MAE) is 0.51 eV, that is to say, the energy error is 47 meV per 1000 atom. The result indicates that our model has good accuracy in the large defective graphene system. With the increase in the number of defects, the network will have a certain deviation in the energy prediction of each defect, and the increase in the concentration will cause the error accumulation of the defect energy. It can be seen from Fig. 3(b) that as the defect concentration of the system increases, the error accumulation also increases, but much less than that of the defect concentration. For example, at 5%, the MAE is 0.51 eV, while at 40%, the MAE is only 1.85 eV. (If the error increases linearly with the defect concentration, it will be 4.08 eV.)

In order to verify the sensitivity of the model to the prediction of the formation energy of a single type of defect, we put different types of defects into our trained CNN model to predict and compare with the DFT results. Among the five different types of defects, the maximum error from the true value is double vacancy (DV) (5555-6-7777) in which the energy difference between CNN and DFT is 0.12 eV, as shown in Fig. 3(c). Therefore, at low defect concentration, the model can still accurately identify the influence of defects and its energy will
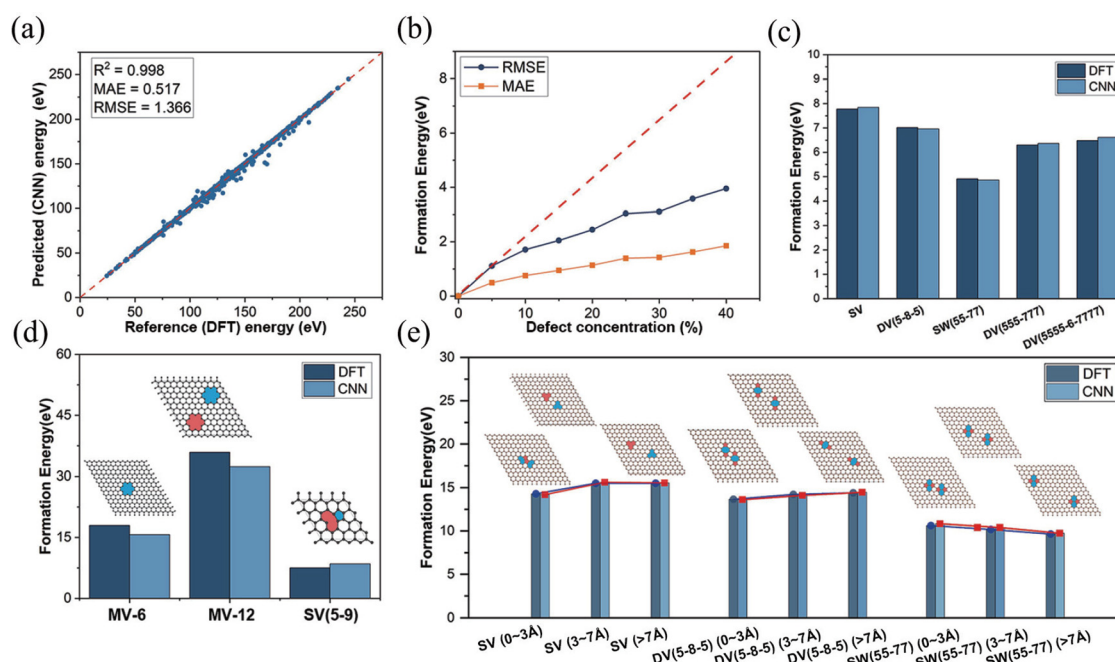
FIG. 3. Evaluation of CNN model. (a) The predictive performance of the trained network for defective graphene. The red dashed line is the ideal fitting line. (b) The prediction results under different defect concentrations. The red dotted line is the error accumulation reference line. (c) The network predicts the results of a single defect of different types. (d) The network's prediction results for several defective structures that do not exist in the data set. MV-6 represents a multi-vacancy defect with six missing atoms, and MV-12 represents a multi-vacancy defect with a total of 12 missing atoms. (e) The prediction results of the network under different operating distances for different combinations of defects.

not be overwhelmed by a large area of perfect graphene background. This illustrates the effectiveness of the convolution and pooling operations of our CNN model.

We further choose single vacancy (SV)–SV, DV (5-8-5)–DV (5-8-5), and SW (55-77)–SW (55-77) to verify the recognition accuracy of the model under different interaction distances. By comparing the formation energy predicted by CNN and DFT, we find that the trends in energy change in these three groups are consistent with our DFT results. For the total energy of any interaction distance under the combination of these two defects, the error predicted by CNN is less than 0.3 eV, as shown in Fig. 3(e). Therefore, the trained CNN model has very good recognition accuracy for the interaction between different defect types at different distances.

In order to test the generalization ability of our CNN model, we select three unknown structures that do not exist in the database: MV-6, MV-12, and SV (5–9) for testing. Clearly, our CNN model gives close values to DFT, as shown in Fig. 3(d). This indicates that the trained CNN learns the potential laws behind the data through the network in the process of deep network learning and obtains the mapping relationship between structural features and energy.

As we all know, the total energy of the defective graphene is not only related to the type and number of defects, but also closely related to the distribution of defects. Therefore, we use above well-trained model to analyze how defect distribution affects the formation energy. The aggregation index is used to describe the distribution of the defect system here (Fig. S11).

For a system that contains only a single type of defect, the formation energy is also different due to different defect distributions. As

shown in Fig. 4(a), when only SV defects are considered in the entire system, as the index of aggregation increases, the energy of the entire system decreases. Since SV defects have dangling bonds, when the distance between SV and surrounding defects is too close, it will cause the dangling bonds of SV to re-form more stable chemical bonds with surrounding atoms due to the interaction, which will largely reduce the formation energy. Furthermore, we consider the existence of two different types with opposite energy changes in the system. It is found that under a certain aggregation index, when the amount ratios of SW–SV are different, the defect formation energy of the entire system will change, as shown in Fig. 4(b). When the system contains only SW, the formation energy is lower than that of the system containing only SV defects, which also complies with our DFT results. When the quantity ratio of SW and SV is 70%–30%, the competition is very fierce, and the energy of the entire system reaches its peak. On both sides of the ratio (SW in 70% and SV in 30%), the energy shows a decreasing trend. Therefore, in the experiment, try to avoid competition between vacancy defects and SW defects, so as to make the entire system more stable. Furthermore, under different aggregation indexes and defect concentrations, the comparison of CNN model and simple superposition of fragments with individual defects is shown in Figs. S7 and S8 in the supplementary material.

To further verify the applicability of our descriptor and model, we extend the multi-layer descriptor to transition metal dichalcogenide (TMD) monolayers as a more general case of 2D materials. The structures of the monolayer TMDs can be seen as different stacking orders of three atomic sublayers (chalcogen–metal–chalcogen),[27] shown in Fig. 5(a). The bond position, bond length (l), polar angle (θ), and
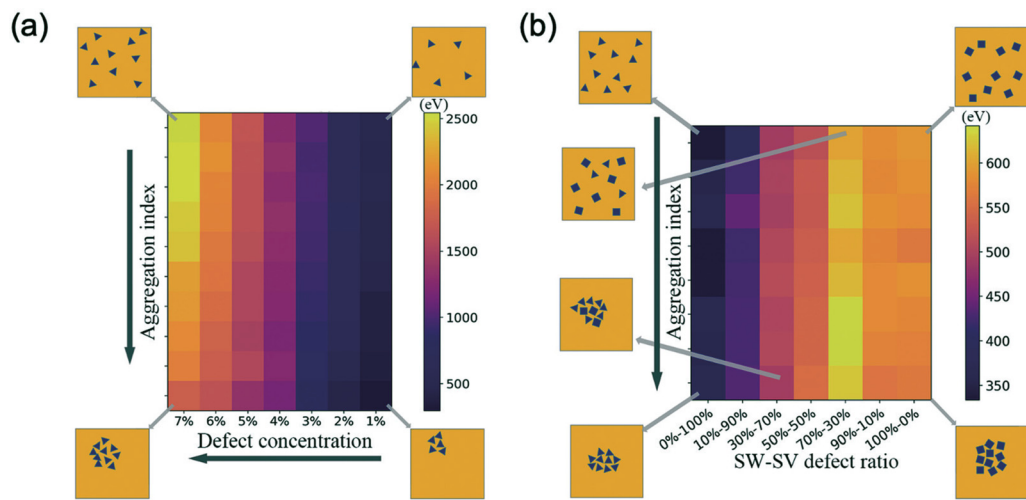
FIG. 4. (a) Phase diagram of aggregation index, defect concentration, and system trap formation energy. Only SV defect is considered here. (b) The phase diagram of different defect quantity ratio and aggregation index and system defect formation energy. There are only two different types of defects, SW and SV, being considered.

azimuthal angle (φ) of 2H-MoS$_2$ are chosen as feature parameters. Thus, structures of MoS$_2$ are encoded into two matrices for each feature parameters corresponding to two layers of Mo–S bonds, as shown in Fig. 5(b).

As shown in Fig. 5(c), our CNN model obtains very accurate prediction results for MoS$_2$. MAE is only 53 meV per 1000 atoms, which is close to graphene results (47 meV per 1000 atoms), proving that our multi-layer descriptor and CNN model can be extended to other
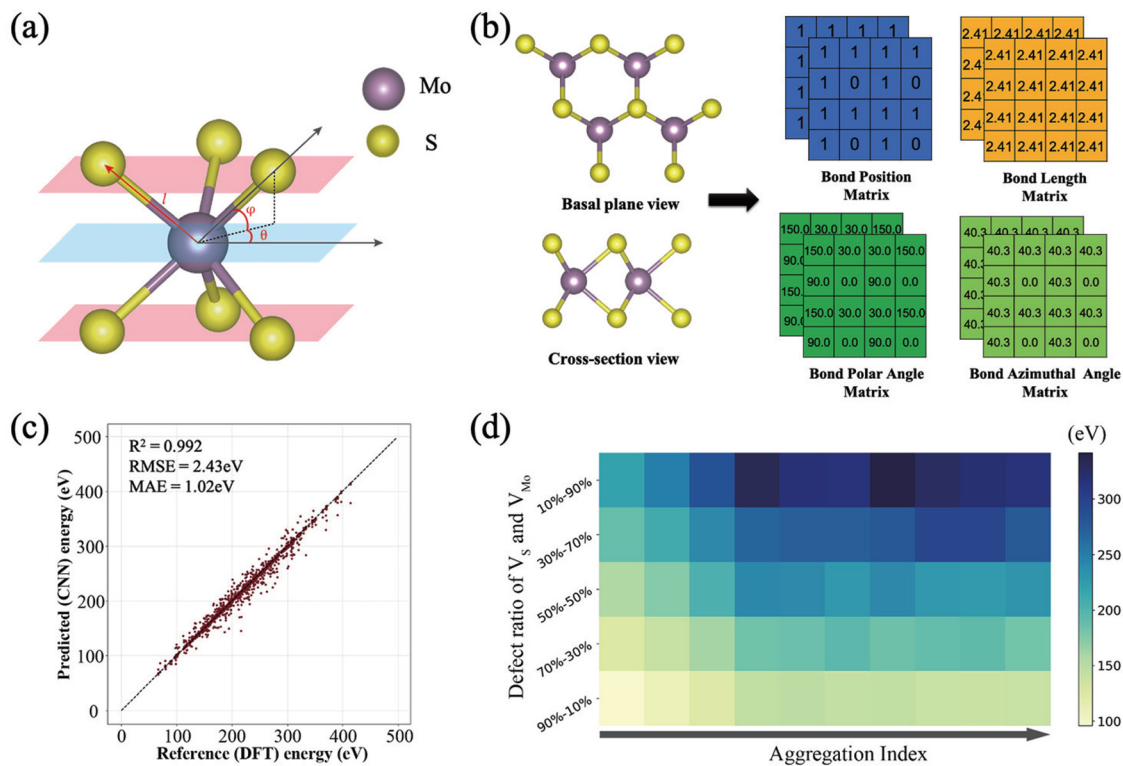

FIG. 5. (a) Schematic diagram of the encoded bond parameters in the structure descriptor of 2H-MoS$_2$. For Mo–S bond, l is the bond length, θ is the polar angle in the plane, and ϕ is the azimuthal angle out of the plane. (b) 2H-MoS$_2$ encoded by multi-layer structure descriptor. (c) The predictive performance of the trained network for defective MoS$_2$. (d) Phase diagrams of formation energies at different aggregation index with different defect ratio of V$_S$ and V$_{Mo}$.

large-scale 2D defect systems. Also, the model shows better performance at high defect concentration, which is similar with graphene results as illustrated in Fig. S14. Furthermore, we use our trained CNN to predict formation energy of defective MoS_2 at different defect ratio and distribution. For simplicity, we only consider the combination of S vacancy (V_S) and Mo vacancy (V_{Mo}). As shown in Fig. 5(d), at the same aggregation, S vacancy is always more stable than Mo vacancy, which is consistent with intuition and DFT results. On the other hand, at the same V_S–V_{Mo} ratio, as the aggregation index reduces, the energy decreases, which follows the same trend as the graphene in Fig. 4. It is worth mentioning that our results show the interaction between S and Mo vacancy is negligible when distance larger than about 9 Å. While for graphene, the cutoff distance is about 15 Å, which indicate that defects aggregation has more influence in graphene that MoS_2.

To summarize, we have developed a multi-layer structure descriptor, which can fully describe the large-scale defective system of 2D materials by encoding the chemical bond information. Based on the deep learning framework, a CNN model that is suitable for our descriptors as input samples was developed to predict the formation energy of defective graphene and MoS_2. After training, CNN can accurately predict the formation energy of large-scale defective graphene (MoS_2) systems at low concentration of defects, and for a graphene (MoS_2) of about 300 nm^2 (650 nm^2), the MAE of the formation energy is less than 47 meV per 1000 atoms (53 meV per 1000 atoms). Based on the trained model, the energy of the large system under different defective distribution can be obtained in DFT accuracy level. Our multi-layer descriptor and CNN network provide a fast and effective method for the study of large-scale defective 2D materials.

See the supplementary material for the DFT calculation results, details of descriptors, and discussion of the CNN model.

AUTHOR DECLARATIONS
Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1]N. M. R. Peres, F. Guinea, and A. H. Castro Neto, "Electronic properties of disordered two-dimensional carbon," Phys. Rev. B **73**, 125411 (2006).

[2]F. Hao, D. Fang, and Z. Xu, "Mechanical and thermal transport properties of graphene with defects," Appl. Phys. Lett. **99**, 041901 (2011).

[3]G. Yang, L. Li, W. B. Lee, and M. C. Ng, "Structure of graphene and its disorders: A review," Sci. Technol. Adv. Mater. **19**, 613–648 (2018).

[4]H. Qiu, T. Xu, Z. Wang, W. Ren, H. Nan, Z. Ni, Q. Chen, S. Yuan, F. Miao, F. Song, G. Long, Y. Shi, L. Sun, J. Wang, and X. Wang, "Hopping transport through defect-induced localized states in molybdenum disulphide," Nat. Commun. **4**, 2642 (2013).

[5]H. Nan, Z. Wang, W. Wang, Z. Liang, Y. Lu, Q. Chen, D. He, P. Tan, F. Miao, X. Wang, J. Wang, and Z. Ni, "Strong photoluminescence enhancement of MoS_2 through defect engineering and oxygen bonding," ACS Nano **8**, 5738–5745 (2014).

[6]S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, and J. Ni, npj Comput. Mater. **5**, 84 (2019).

[7]P. M. Tagade, S. P. Adiga, S. Pandian, M. S. Park, K. S. Hariharan, and S. M. Kolake, "Attribute driven inverse materials design using deep learning bayesian framework," npj Comput. Mater. **5**, 127 (2019).

[8]N. Sun, J. Yi, P. Zhang, H. Shen, and H. Zhai, "Deep learning topological invariants of band insulators," Phys. Rev. B **98**, 085402 (2018).

[9]S. Feng, H. Fu, H. Zhou, Y. Wu, Z. Lu, and H. Dong, npj Comput. Mater. **7**, 10 (2021).

[10]A. Cecen, H. Dai, Y. C. Yabansu, S. R. Kalidindi, and L. Song, "Material structure-property linkages using three-dimensional convolutional neural networks," Acta Mater. **146**, 76–84 (2018).

[11]S. Sanyal, J. Balachandran, N. Yadati, A. Kumar, P. Rajagopalan, S. Sanyal, and P. Talukdar, "MT-CGCNN: Integrating crystal graph convolutional neural network with multitask learning for material property prediction," arXiv:1811.05660 (2018).

[12]D. R. Cassar, A. C. P. L. F. de Carvalho, and E. D. Zanotto, "Predicting glass transition temperatures using neural networks," Acta Mater. **159**, 249–256 (2018).

[13]X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, "A transfer learning approach for microstructure reconstruction and structure-property predictions," Sci. Rep. **8**, 13461 (2018).

[14]Y. Song, E. M. D. Siriwardane, Y. Zhao, and J. Hu, "Computational discovery of new 2D materials using deep learning generative models," ACS Appl. Mater. Interfaces **13**, 53303–53313 (2021).

[15]Z. Wang, Q. Wang, Y. Han, Y. Ma, H. Zhao, A. Nowak, and J. Li, "Deep learning for ultra-fast and high precision screening of energy materials," Energy Storage Mater. **39**, 45–53 (2021).

[16]D. Jha, L. Ward, A. Paul, W. K. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "Elemnet: Deep learning the chemistry of materials from only elemental composition," Sci. Rep. **8**, 17593 (2018).

[17]N. C. Frey, D. Akinwande, D. Jariwala, and V. B. Shenoy, "Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing," ACS Nano **14**, 13406–13417 (2020).

[18]W. Ye, C. Chen, Z. Wang, I. H. Chu, and S. P. Ong, "Deep neural networks for accurate predictions of crystal stability," Nat. Commun. **9**, 3800 (2018).

[19]J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," Chem. Mater. **29**, 5090–5103 (2017).

[20]T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).

[21]P. Hundi and R. Shahsavari, "Deep learning to speed up the development of structure-property relations for hexagonal boron nitride and graphene," Small **15**, 1900656 (2019).

[22]J. Paier, R. Hirschl, M. Marsman, and G. Kresse, "The Perdew–Burke–Ernzerhof exchange-correlation functional applied to the G2-1 test set using a plane-wave basis set," J. Chem. Phys. **122**, 234102 (2005).

[23]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies," npj Comput. Mater. **1**, 15010 (2015).

[24]S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen,

J. Gath, K. W. Jacobsen, J. Jørgen Mortensen, T. Olsen, and K. S. Thygesen, "The computational 2D materials database: High-throughput modeling and discovery of atomically thin crystals," 2D Mater. **5**, 042002 (2018).

[25]L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "Dscribe: Library of descriptors for machine learning in materials science," Commun. Comput. Phys. **247**, 106949 (2020).

[26]F. Banhart, J. Kotakoski, and A. V. Krasheninnikov, "Structural defects in graphene," ACS Nano **5**, 26–41 (2011).

[27]S. Manzeli, D. Ovchinnikov, D. Pasquier, O. V. Yazyev, and A. Kis, "2D transition metal dichalcogenides," Nat. Rev. Mater. **2**, 17033 (2017).