Prediction Of Material Properties By Neural Network Fusing The Atomic Local Environment And Global Description: Applied To Organic Molecules And Crystals

Deyu Xia[1,a], Ning Li[1,b*], Pengju Ren[2] and Xiaodong Wen[2]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing, China
[2]Synfuels China Technology Co., Ltd., Beijing, China

Abstract. Machine learning has brought great convenience to material property prediction. However, most existing models can only predict properties of molecules or crystals with specific size, and usually only local atomic environment or molecular global descriptor representation be used as the characteristics of the model, resulting in poor model versatility and cannot be applied to multiple systems. We propose a method that combines the description of the local atomic environment and the overall structure of the molecule, a fusion model consisting of a graph convolutional neural network and a fully connected neural network is used to predict the properties of molecules or crystals, and successfully applied to QM9 organic molecules and semiconductor crystal materials. Our method is not limited to a specific size of a molecule or a crystal structure. According to the calculation principle of the properties of the material molecules, the influences of the local atomic environment and the overall structure of the molecules on the properties are respectively considered, an appropriate weighting ratio is selected to predict the properties. As a result, the prediction performance has been greatly improved. In fact, the proposed method is not limited to organic molecules and crystals and is also applicable to other structures, such as clusters.

1 INTRODUCTION

Studies at the molecular and atomic levels play an important role in better understanding the properties of materials. However, the traditional quantum mechanics method is inefficient in calculating molecular properties, which limits the development of materials science to a certain extent[1]. Therefore, the search for an alternative to density functional theory (DFT) has become an active hot research topic[2][3]. In recent years, machine learning methods have been attracted lots of attention and have been applied to calculate material properties in specific fields or for general purposes. It has made outstanding achievements such as crystal structure prediction[4]. In particular, many researchers have developed a machine learning model based on neural network[5][6] or ridge regression[7][8] to predict the various properties of different molecular systems, including crystals, organic molecules, clusters, and so on[9-17]. However, systematically improving the accuracy and generalization ability of machine learning models for different systems remains a major challenge for scientists in both the computer field and application fields of chemistry and materials.

In material prediction, any successful machine learning method relies on the input features, whose role is to establish a statistical relationship between the molecular structure and properties of materials[18]. High-quality features not only contain rich molecular structure information, can accelerate the training process, and improve the accuracy of machine learning model. For the application of machine learning model for material property prediction, descriptors satisfying the rotational invariance of molecular or crystal structure are generally selected for feature construction[19][20]. The original structural representation of a molecule or crystal is constructed from descriptors, and the dimensions usually depend on the size of the system. For disordered systems, applying the features constructed by descriptors directly to machine learning models will result in high training costs and poor results. The features constructed by descriptors represent either the local atomic environment or the global molecular environment[21][22]. In the previous studies, usually single structure representation is used for ML process. Although specific prediction tasks can be accomplished, the accuracy is not enough. For molecules and materials, some properties mainly depend on the local structure. Such properties have atomic additivity, such as the formation energy of the system. Therefore, a new method is proposed in this paper. When predicting the molecular properties of materials, it not only uses the atomic local environment representation, but also constructs a new molecular global structure representation based on the atomic environment representation. The two are combined and the neural network is used to learn and predict the various properties of material molecules. Our

[a]deyu_xia@163.com
[*]Corresponding author: [b]ningli.ok@163.com

method makes full use of the characteristics of atomic local environment and molecular global environment, not only does not limit the size of the molecule, but also can achieve ideal results in different systems.

Our method can be extended to different systems and various properties can be calculated. We have successfully applied the method presented in this paper to different systems (organic molecules and transparent conductors) and compared them against well-known machine learning prediction algorithms. In the following section, we discuss the method of property prediction in this paper in detail.

2 METHOD

Similar to other machine learning methods for calculating potential energy, the steps of our method are shown in Figure 1. First, we need to select an appropriate molecule to calculate its physical property. After data screening and pre-processing, training set and test set are obtained. Please note that supercell processing is carried out on some crystals, for the structural particularity of crystals, which will be described in detail later. We then analyze and construct the molecular features, including selecting descriptors and setting reasonable parameters. In order to verify the validity and rationality of features, feature correlation analysis is needed. Next, we should design a reasonable fitting model structure and conduct model training. Finally, the model can be applied after verification.

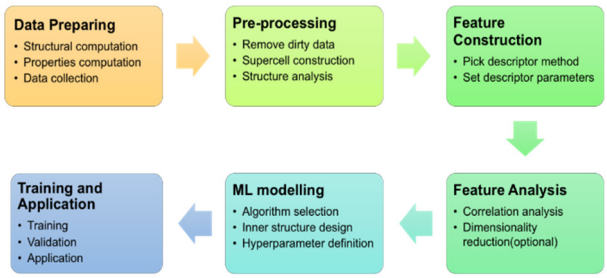

Fig. 1 Steps to predict material properties using machine learning methods.

In order to satisfy the invariance of atomic rotation and translation, Smooth Overlap of Atomic Positions (SOAP) is introduced as the descriptor. The SOAP formula is given below:

$$P(r)_{nn'l}^{Z_1Z_2} = \pi\sqrt{\frac{8}{2l+1}}\sum_{m} c_{nlm}^{Z_1}(r)^* c_{n'lm}^{Z_2}(r) \qquad (1)$$

where n and n' is the index of different radial basis function, from 0 to n_{max}; l is the angle of the harmonic function, from 0 to l_{max}; Z_1 and Z_2 is atomic species. c_{nlm}^{Z} is calculated using the following inner product:

$$c_{nlm}^{Z}(\mathrm{r}) = \iiint_{R^3} dV\, g_n(\mathrm{r}) Y_{lm}(\theta,\phi)\rho^{Z}(\mathrm{r}) \qquad (2)$$

where $\boldsymbol{r}$ is a location in space; $\rho^{Z}(\mathrm{r})$ is the Gaussian smoothing atomic density of atom Z; $Y_{lm}(\theta,\phi)$ is a harmonic function; and $g_n(\mathrm{r})$ is the radial basis function.

With the SOAP method, we can get a local atomic environment representation of each atom, combined with the graph structure, we can get the contribution of each local atom environment to the whole molecule or crystal.

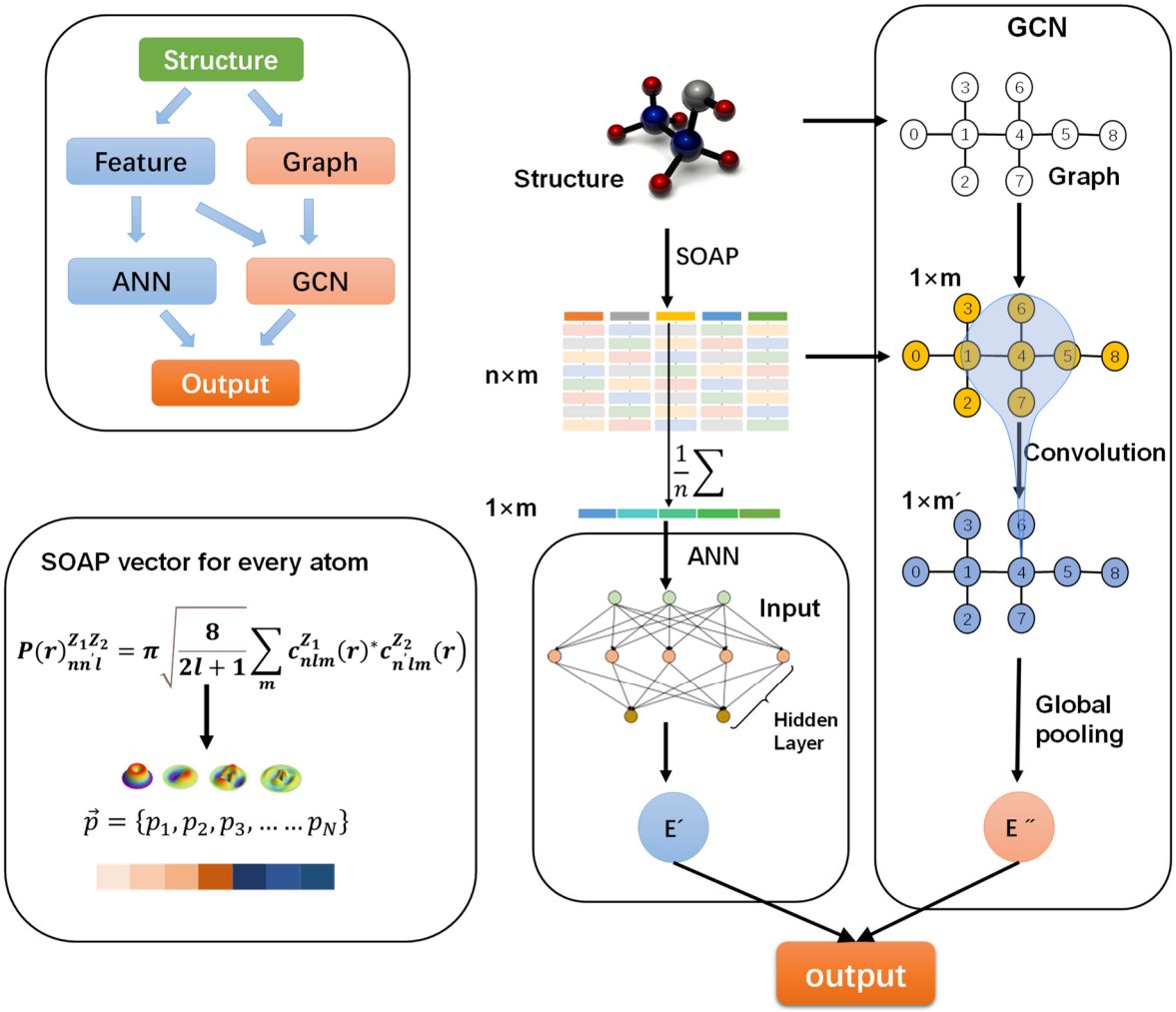

Fig. 2 Explains how models make property predictions.

Figure 2 shows the main concept of the proposed method: the structure of a molecule or crystal is represented by a graph in which each point represents an atom and the edges represent the existence of associated bonds between the two atoms. A convolution layer is added to the graph to obtain the energy contribution of each atom. An ANN is integrated into the model to obtain an additional energy contribution. The two energy contributions are combined to yield the energy of the whole molecule or crystal.

First, a SOAP matrix is generated for each molecular structure. The soap matrix is processed in two ways. 1)MALE: The column vectors of each SOAP matrix are averaged to obtain an eigenvector with dimensions of 1*m, which is used to describe a molecular structure. 2)AME: Each atom and its associated atoms are traversed to construct an atomic graph of each molecular structure. Each node in the topological graph represents an atom in the molecule, and the feature vector of the point is still represented by the SOAP vector of the atom. Based on AME and MALE methods, the molecular topology graph and the global environment representation of the molecule were obtained, and then the properties of the two parts were extracted by combining graph convolutional neural network and fully connected neural network. The final properties of the molecule are predicted as a weighted sum based on different tasks and different systems. 1) GCN: The graph structure and atomic feature representation generated by the AME method are used as input, and the features are updated by graph convolution through message passing. Each vertex represents an atom in the molecule, and each edge represents a bond between two atoms. Note that this graph is an undirected graph. After the pooling layer, the contribution of all the atoms to the molecular properties was obtained. 2) ANN: The atomic structure representation obtained by the MALE method is used as the input to the ANN network. The contribution of property value is calculated through the full connection layer. After GCN and ANN, two parts about the contribution value of the property are obtained, and the weighted sum of them is used to get the final property value as the output.

In the investigated model, the GCN contains two convolutional layers and a pooling layer. The formula of the convolution layer is as follows:

$$h_i^{(l+1)} = \sigma\left(b^l + \sum_{j\in N(j)} \frac{1}{c_{ij}} h_i^l W^l\right) \quad (3)$$

where $N(j)$ is the set of neighbors of node i; c_{ij} is the product of the square root of node degrees($c_{ij} = \sqrt{|N(i)|}\sqrt{|N(j)|}$); and σ is an activation function.

The formula of the global pooling layer is as follows:

$$r^{(i)} = \sum_{k=1}^{N_i} x_k^i \quad (4)$$

Convolution and pooling have been used very successfully in image processing, natural language processing and other fields. Graphic convolution has also been used in molecular modeling.

In our method, the vector of the atomic local environment is processed by GCN, and through the message passing mechanism of a graph neural network, each node in the graph updates its own feature vector, which can also be called the hidden state of the graph. After two convolution operations, the contribution value of each atom to the local properties can be obtained, and the overall molecular property value can be calculated by pooling operation. MALE feature vectors are extracted by ANN, which is the only representation to describe the structure of a single molecule. The properties of molecular structures can be calculated by using fully connected neural networks. The properties predicted by the two parts of the neural network will determine the properties of molecular structure together.

3 RESULTS AND DISCUSSION

The proposed method can mainly be used to predict various properties of molecules in disordered systems; thus, we apply this method to two different systems: organic molecules, for which the UO and GAP are calculated; and crystals, for which the formal and bandgap energies are computed. The construction and selection of features in machine learning are extremely important, and the correlation between features should be considered. The proposed method was used to construct a total of 234 features for each atom, where the correlation degree (the Pearson correlation coefficient) among the features is shown in Fig. 3.a. The distribution statistical graph of the characteristic correlation degree is shown in Fig. 3.b. It can be seen that more than 50% of the features have a correlation degree between 0 and 0.1, and more than 90% of the features have a correlation degree below 0.5, indicating a very low correlation between features. At the same time, more features can contain more information, which is helpful for machine learning fitting model.

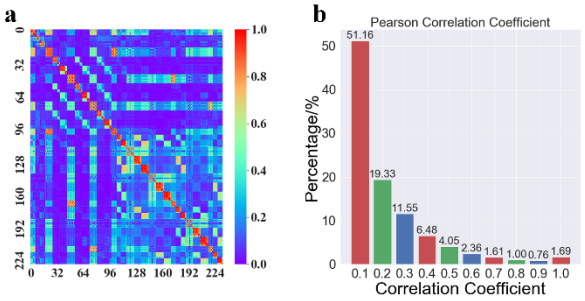

Fig. 3 Plot of the Pearson correlation coefficient (absolute values) between the features.

The combination of the enhanced model structure and information-rich features developed in this study should theoretically produce improved model fitting results. The abovementioned method was used to successfully predict the U0 energy and GAP in organic molecules in the QM9 data set. To consider the different contributions of the local atomic and global molecular environments to the whole mass, the errors in the predicted U0 and GAP energies were calculated using different GCN:ANN ratios. The results are shown in Fig. 4. The most accurate predictions for the U0 and GAP energies are obtained using a GCN:ANN ratio of 0.5.

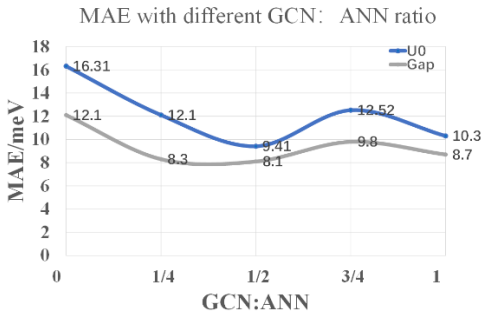

Fig. 4 Plot the error results of different proportions of GCN and ANN

We also compared with the famous property prediction model[23] of organic molecules, the comparison results are shown in Figure 5. It can be found that when our model predicts U0 and Gap, the MAE is the smallest, indicating that our model has the best prediction effect. Because our model can make full use of atomic local information and molecular global representation, it has great advantages over a single feature representation.

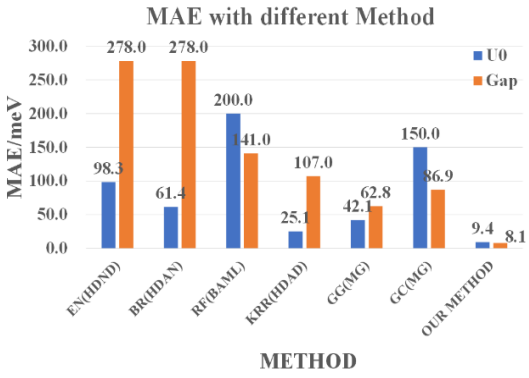

Fig. 5 Plot the MAE with different Method

We also collected 2,400 pieces of crystal data to analyze their structural properties. Crystal structure is composed of $(Al_xGa_yIn_z)_2O_3$, as shown in Figure 6.a, most of the crystals are composed of Ga-Al-In-O. Note that in order to reasonably calculate the crystal diagram, we performed supercell processing on the original crystal structure, and the statistics of the number of atoms contained in the crystal after the supercell are shown in Figure 6.b. The height in the three crystal directions is compared with the length of the maximum covalent chemical bond to determine whether supercellular processing of the crystal structure is required. If the height in a given direction is less than 2 times the covalent chemical bond length, supercellular processing should be performed in that direction.

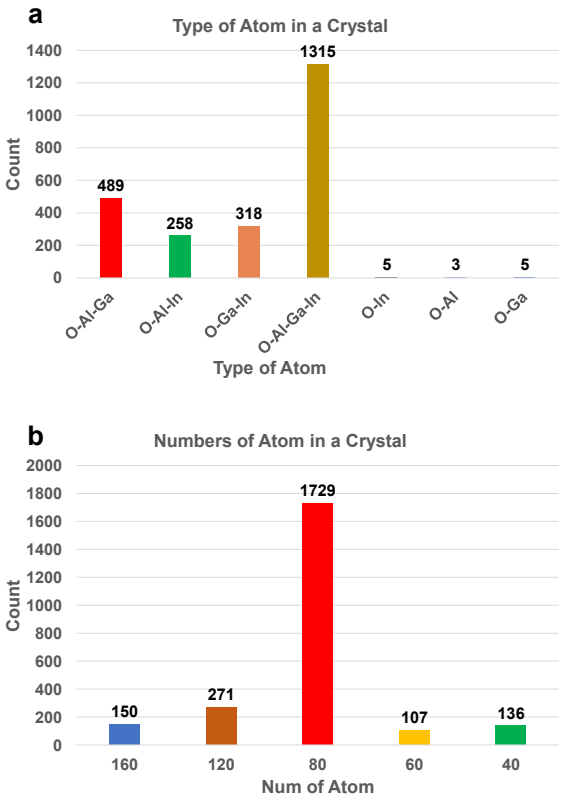

Fig. 6 The components of the $(Al_xGa_yIn_z)_2O_3$ dataset. a): The number and types of atoms contained in each structure in the dataset. b): The number of atoms contained in each structure after supercell in the dataset.

As shown in Fig. 7.a, the abscissa is the predicted value of the model, and the ordinate is the actual energy value. It can be seen that points in both the training set and the test set are distributed near the line y=x, indicating that our fitting performance is good. The error statistical figure of each crystal predicted value is shown as Fig. 7.b. We can see that the model in the training set has a good performance, and most errors are lower than 20 meV/atom, indicating that our model has a good fitting performance and has reached the best state during training. In the test set, the prediction error of more than 84% crystal structure is less than 20 meV /atom, and only 1.24% crystal prediction error is more than 100 meV /atom, which indicates that our model has good robustness. Notably, our model has a MAE of 12.08 meV/atom on the test set, which is a significant improvement over other methods on the same dataset. This is because the feature vector performs better when the feature correlation degree is low, and our model makes full use of and mines the information between features, so a relatively good result we got.

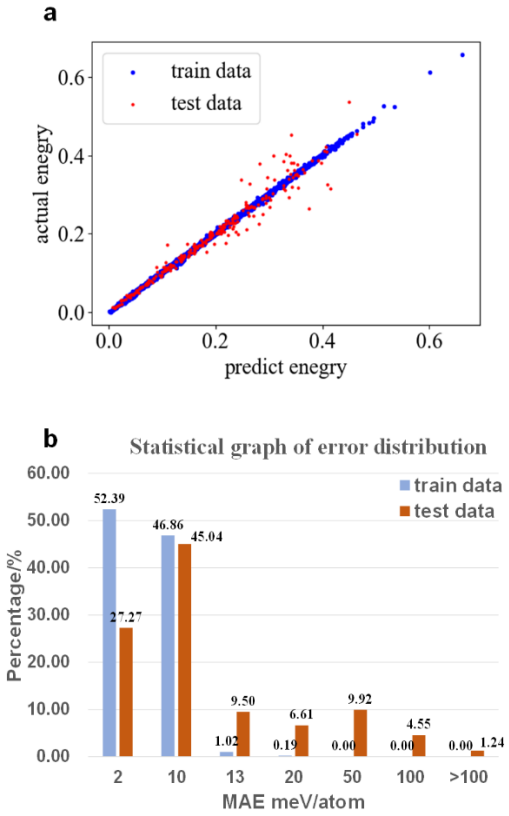

Fig. 7 Performance of GCN and ANN fusion model. a) 2D histogram showing the predicted formation energy per atom compared to the DFT calculated value. b) Histogram of distribution of error in energy. Note that the mean error in the test set was 12.08 meV/atom, more than 94% of the structural error was less than 50 meV/atom, and only %1.24 of the structural prediction error exceeded 100meV.

We have also compared with the best calculation method of crystal potential energy listed in Literature[24], and the comparison results are shown in Table 1. We can find that when calculating Formation energy, the MAE calculated by our method is the smallest, indicating that our method performs best. In the calculation of Bandgap, the error is large, and there is still a gap with the best method.

Table 1 Comparison of the predictions of the proposed method and other models using the same data set

METHOD	# of data	MAE	
		Formation energy(meV/atom)	Bandgap(meV)
SIFF+GDB	2400	18.5	117.0
CGCNN+GDB	2400	18.7	**115.0**
GCN	2393	13.0	138.0
GCN+ANN*	2393	**12.08**	143.8

4 CONCLUSIONS

In summary, reasonable features are constructed and selected to describe the molecular structure of materials, and reasonable models are constructed to fully mine the information between the features, thereby improving the

predicted molecular properties and significantly reducing the fitting error. A lower MAE is obtained using the proposed method compared to those of the most notable methods developed over the past few years for the same dataset, indicating that the proposed method outperforms existing methods. The performance of the GCN and neural network fusion model is also higher than that of the GCN model. This result is obtained because the contribution degrees of local atomic and global molecular features to the molecular attribute values are fully considered in the proposed model, and feature processing methods are manually incorporated to make the model more effective. Note that the proposed method is not limited to certain molecular structures or sizes because the size of the molecular graph is not restricted. Thus, the properties of molecular structures of any size can be predicted (if data are available). The proposed method is also not limited to crystals and organic molecules. Property prediction models can also be built for other systems, such as clusters, using molecular graph structures, and the proposed method could be used to perform feature extraction and property prediction. We will apply the proposed method to other systems in a future study.

ACKNOWLEDGMENTS

Deyu Xia, Ning Li, Pengju Ren and Xiaodong Wen acknowledge funding from the Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Industry-University Cooperation Base between Beijing Information S&T University and Synfuels China Co., Beijing 100101, China.

REFERENCE

1. Ulissi, Z. W., Medford, A. J., Bligaard, T., Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. Nature communications, **8**,1, 1-7. (2017).
2. Faber, F. A., Hutchison, L., Huang, B., et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. Journal of chemical theory and computation, **13**,11, 5255-5264. (2017).
3. Shen, L., & Yang, W. Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks. Journal of chemical theory and computation, **14**,3, 1442-1455. (2018).
4. Rosenbrock, C. W., Homer, E. R., Csányi, G., Hart, G. L. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. NPJ Computational Materials, **3**,1, 1-7. (2017).
5. Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. Journal of Physics: Condensed Matter, **26**,18, 183001. (2014).
6. Wong, S. Y., Bund, R. K., Connelly, R. K., et al. Modeling the crystallization kinetic rates of lactose via artificial neural network. Crystal growth & design, **10**,6, 2620-2628. (2010).
7. Bartók, A. P., Csányi, G. G aussian approximation potentials: A brief tutorial introduction. International Journal of Quantum Chemistry, **115**,16, 1051-1057. (2015).
8. Bartók, A. P., Payne, M. C., Kondor, R., et al. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters, **104**,13, 136403. (2010).
9. Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., et al. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. The journal of physical chemistry letters, **6**,12, 2326-2331. (2015).
10. Schütt, K. T., Arbabzadah, F., Chmiela, S., et al. Quantum-chemical insights from deep tensor neural networks. Nature communications, **8,**1, 1-8. (2017).
11. Yao, K., Herr, J. E., Toth, D. W., et al. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. Chemical science, **9**,8, 2261-2269. (2018).
12. Huang, S. D., Shang, C., Kang, P. L., et al. Atomic structure of boron resolved using machine learning and global sampling. Chemical science, **9**,46,, 8644-8655. (2018).
13. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, **4**,2, 268-276. (2018).
14. Ramprasad, R., Batra, R., Pilania, G., et al. Machine learning in materials informatics: recent applications and prospects. npj Computational Materials, **3**,1, 1-13. (2017).
15. Shapeev, A. V. Applications of machine learning for representing interatomic interactions. *In Computational Materials Discovery*. Royal Society of Chemistry. (2018).
16. Behler, J. Perspective: Machine learning potentials for atomistic simulations. The Journal of chemical physics, **145**,17, 170901. (2016).
17. Xie, T., Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters, **120**,14, 145301.1-145301.6. (2018).
18. Khorshidi, A., Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. Computer Physics Communications, **207**, 310-324. (2016).
19. Bartók, A. P., Kondor, R., Csányi, G. On representing chemical environments. Physical Review B, **87**,18, 184115. (2013).
20. Imbalzano, G., Anelli, A., Giofré, D., et al. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. The Journal of chemical physics, **148**,24, 241730. (2018).

21. Zhang, K., Yin, L., Liu, G. Physically inspired atom-centered symmetry functions for the construction of high dimensional neural network potential energy surfaces. Computational Materials Science, **186**, 110071. (2021).
22. Huo, H., Rupp, M. Unified representation of molecules and crystals for machine learning. arXiv preprint arXiv:1704.06439. (2017).
23. Kim, H., Park, J. Y., Choi, S. Energy refinement and analysis of structures in the QM9 database via a highly accurate quantum chemical method. Scientific data, **6**,1, 1-8. (2019).
24. Zeledon, J. A. H., Romero, A. H., Ren, P., et al. The structural information filtered features (SIFF) potential: Maximizing information stored in machine-learning descriptors for materials prediction. Journal of Applied Physics, **127**,21, 215108. (2020).