Research on Application of Machine Learning Technology in New Material System

Jiaoyu Wang* and Daxiu Zhang
Bengbu University, Bengbu 233000, China

*Corresponding author: jiaoyuwang@bbc.edu.cn

Abstract. Materials are not only the foundation of the national economy, but also the carrier of high-tech. Machine learning combined with computer science, database theory, statistics, computational mathematics and engineering cannot only show faster calculation speed and reliable predictive ability, significantly improve the efficiency of material calculations, and it can also effectively deal with some systems and problems that are difficult to use traditional simulation computing methods. This article will briefly outline the basic principles of machine learning, introduce several typical algorithms in machine learning models and how machine learning is the application progress in the research of new materials, and the prospects for the future development of machine learning in the field of materials science.

Keywords: New materials, machine learning, material calculation simulation.

1. Introduction
In recent years, machine learning has received widespread attention, especially the artificial intelligence AlphaGo developed by Google has frequently won the human-computer battle with the world champion of Go, showing the great potential of machine learning in people's field of vision. At present, the most widely used is density functional theory. However, there is still a certain gap between the current theoretical calculation results and practical applications. For example, density functional theory simulates calculations under high temperature, high pressure and strong magnetic fields. The results are unsatisfactory [1]. The use of machine learning methods can greatly save calculation time, which can be applied to more complex systems or longer time scales. Therefore, combining machine learning with existing calculation methods will enable new material design It is more efficient and accurate, so as to provide more effective guidance for experiments.

2. Machine learning research methods
The material data set is generated through the material structure check technology and the component replacement strategy, including the training set and the test set. The band gap values of all the material systems in the training set are calculated using the VASP package based on the first principles, and on the training, set Learn to obtain relevant machine learning models, and then use a series of models to predict the unknown band gap value of the new material system in the test set, and then use first-principles high-throughput calculations for verification, and finally obtain a relatively accurate model.
for the predicted band gap [2]. As shown in Figure 1. As shown, it is the framework of the band gap prediction model for the replacement products of amorphous materials.

**Figure 1.** The framework of the band gap prediction model for the sub-replacement products of amorphous materials

2.1. Data set and feature descriptor
In this paper, using 87 non-repetitive amorphous material band gap data, through the material structure check with the existing data in MIP, 356 new non-repetitive structure material systems are obtained, and 32 new materials are randomly selected. The system performs high-throughput calculations, and together with the original 87-like amorphous material data, constitutes a set of 119 training samples used to build the band gap prediction model in this study, and the remaining 324 new systems are used as the test set. When performing material first-principles calculations, the VASP program package based on density functional theory is used. All calculations use the affixed plane wave (PAW) method to process the electronic wave function, and the newly developed SCAN meta-GGA is used for Qualcomm Volume calculation.

2.2. Component replacement
There are many ways to discover new material systems. The common method is to use elements with similar chemical properties to replace the corresponding objects in the crystal structure of known compounds to generate new material systems, and then proceed to the next step of preparation and characterization, computational simulation, and property research [3]. For example, when the first
superconducting phosphorus oxide LaFeAsO1-xFx was discovered, crystal chemists began to use other rare earth elements (such as samarium instead of lanthanum) to synthesize many other isostructural new compounds, similar to those the component substitution method has certain practical value for large-scale search of new inorganic compounds.

2.3. Determination of evaluation indicators
This article uses the mean absolute error (MAE), mean square error (MSE) and coefficient of determination ($R^2$) to observe and measure the prediction accuracy of each model on the test set, and compare the performance differences between different models. In short, MAE The smaller the MSE and the closer to 0, the larger the $R^2$ and the closer to 1, the better the performance of the model and the higher the prediction accuracy. The following are several types of error solving calculation models.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

(1)

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

(2)

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}$$

(3)

In formulas (1)–(3), $y_i$ is the high-throughput calculation value of the band gap of 50 amorphous materials randomly selected from the test set; $\bar{y}_i$ is the average value; $\hat{y}_i$ is the predicted value of the regression model corresponding to $y_i$, $i=1,2,...,N$, $N=50$. In order to obtain the true value of the band gap and the three index values of MAE, MSE and $R^2$ of the model prediction result, to avoid random errors, a 10-fold cross-check is used to integrate Determine the predictive ability of each regression model. In the k fold (k value in this article is 10) cross-check, the original data set is randomly divided into k sub-sample sets of equal size [4]. In the k sub-sample set, keep 1 sub-sample set as the verification data of the test model, the remaining k-1 sub-sample sets are used to train the model. Then the cross-validation process is iterated k times, in which each sub-sample set is used as the verification data only once. Finally, it will be from each time the cross-validation prediction results are averaged to get an overall estimated performance index value.

3. Application of machine learning algorithms in the field of new materials

3.1. Predict the properties of materials

3.1.1. Research summary. At present, the design and development of materials mainly rely on traditional experimental methods to synthesize and test samples. The development process of new materials is time-consuming and low in efficiency. Theoretical simulation calculations, especially the introduction of high-throughput screening technologies, have accelerated the development of materials. Research and development process. Compared with the current mainstream theoretical calculation methods, machine learning has fast calculation speed, high flexibility, strong generalization ability, and has the potential to be combined with experiments, which can effectively speed up the design and development process of materials.

In the research of amorphous materials, machine learning has also made some progress. For glassy substances, there is only a weak correlation between its structure and dynamic properties. Some
experimental groups use machine learning methods to characterize glassy materials. The "softness" of its local structure, and explored the correlation between its dynamic properties and "softness".

Similarly, in disordered solid materials, the use of machine learning methods can also effectively study the flow of defects and the rearrangement of particles, so as to summarize the universal microstructure characteristics of disordered materials. Material synthesis and processing, Especially the reaction mechanism of some complex compounds, at present, there is still a lack of deep understanding and cognition [5]. The research and development of materials mainly rely on exploratory attempts and synthesis, and the accumulation of experience is used to determine a reasonable synthesis and processing process. As a data-driven calculation method, learning can skillfully combine experimental data and provide a potential method for process optimization.

3.1.2. Case description. Take the simulation of the forming process of a large steel ingot (about 20t) as an example. Simultaneously pour three shapes of solid steel ingot, hollow steel ingot and liquid core extruded ingot. After cooling, take out the steel ingot and compare the temperature and interface movement of these three ingots process, surface stress, residual stress, porosity, segregation and other differences, and then use these parameters to optimize the process [6]. The simulated material is 45# steel ingot, and its parameters are as follows: initial temperature is 1823.15K, Poisson's ratio is 0.269, elasticity The modulus is 210GPa, the density is 7833kg/m³, the liquid heat capacity is 580.61J/(kg·K), the solid heat capacity is 465J/(kg·K), the solid state thermal conductivity is 319.52W/(m·K), liquid thermal conductivity is 309.05W/(m·K), the solidification temperature is 1773.15K, the mold heat transfer coefficient is 320.7W/(m²·K), the solidification latent heat is 205kJ/kg, and the liquidus temperature is 1768.15K. There are currently two solidification methods, directional solidification and volume solidification. This simulation belongs to volume solidification (solidification in all directions).

\[ q = \left( R_c \rho \Delta h - c \rho \varepsilon \right) \frac{V}{A} \]  (4)

Where: \( q \) is the interface heat flux density (J/(m²·s)); \( R_c \) is the volume solidification rate (m/s); \( \rho \) is the average density (kg/m³); \( \Delta h \) is the latent heat of solidification (kJ/kg); \( c \) is the heat capacity (J/K); \( \varepsilon \) is the cooling rate (K/s); \( V \) is the ingot volume (m³); \( A \) is the ingot interface area (m²). The solidification process of 45# steel at high temperature changes from the liquid phase to the solid phase, so the phase change heat transfer needs to be considered, and the heat transfer during the phase change process is deduced from equation (5).

\[ \rho C_p \frac{\partial T_s}{\partial t} + \rho C_p u \nabla T_s + \nabla q = Q_{\text{total}} \]  (5)

The calculation of \( q \) and \( \rho \) in formula (5) is shown in formula (6) and formula (7).

\[ q = -k \nabla T_s \]  (6)

\[ \rho = \theta \rho_{\text{phase1}} + (1 - \theta) \rho_{\text{phase2}} \]  (7)

Where: \( C_p \) is the heat capacity at normal pressure (J/(kg·K)); \( k \) is the thermal conductivity; \( T_s \) is the temperature (K); \( t \) is the time (s); \( u \) is the velocity field (m/s); \( \theta \) is the solid-liquid ratio; \( \nabla \) means that the derivative of the function in each orthogonal direction is multiplied by the unit vector in each direction.
After completing the equation setting and grid setting, set the calculation process of the steel ingot, set the time range to the process of solidification to 1000s, and do iterative calculation with each step of 0.1s. The simulation content is that during the forming process of the 45# steel ingot, the molten steel is first poured into the mold and after it is cooled to solidification, the steel ingot is taken out for analysis. The diameter of the upper base of the steel ingot is 1358mm, the diameter of the lower base is 1158mm, the height is 2000mm, and about 20t. Its shape and size are shown in Figure 2.

![Figure 2. Solid steel ingot](image)

The COMSOL simulation results are shown in Figure 2. The heat is dissipated from the outer wall. Since the pouring is poured by the small side (1158mm diameter side), the small side (1158mm diameter side) is set to continuously provide heat source. Figure 2 shows the temperature longitudinal section after solidification for 2.9 hours.

![Figure 3. Solid steel ingot temperature (K)](image)

3.2. Improved material theoretical simulation calculation method

There are many ways to combine density functional theory with machine learning. The simplest and most direct method is to use the calculation results of density functional theory as training data for machine learning for learning and prediction. This idea has some examples. The Cedar experimental group used a combination of machine learning and DFT methods to find ternary oxygen-containing compounds that do not exist in nature. The Tanaka experimental group combined machine learning and first-principles calculations to design lithium-ion conductor materials. Their experiments showed that, when the consistency between theoretical calculation data and experimental data is guaranteed, machine learning can use experimental data and theoretical calculation data at the same time to perform simulation calculations more efficiently.
Machine learning can not only use the calculation results of the DFT method to train the prediction properties, but also to optimize the calculation process of the DFT method itself. Due to the kinetic energy part of the energy functional in the density functional theory, especially the interacting electrons. The kinetic energy term of the system is unknown, so the Kohn-Sham equation needs to be solved, and the Kohn-Sham equation incorporates all the complexity of the interacting electronic system into the exchange correlation interaction functional, which is an approximation of the exchange correlation functional. The processing determines the accuracy of the DFT calculation, and this part of the approximate calculation is also the most time-consuming part of the DFT calculation. The existing approximate calculation methods have certain limitations, such as high temperature, high pressure, and strong magnetic fields. It is difficult to fit correct and reasonable results [7]. Combine machine learning methods with density functional theory, use machine learning methods to improve the existing DFT-based first-principles calculation methods, and use the flexibility and efficiency of machine learning. It can reduce the amount of calculation to a certain extent and speed up the calculation.

3.3. Combine with high-throughput screening methods

Machine learning and high-throughput screening also have some interoperability. Combining high-throughput screening with machine learning may further improve the efficiency of theoretical simulation calculations. High-throughput screening methods can simultaneously analyse thousands of possible possibilities in a short period of time. The materials are filtered to find a system that meets specific performance standards for experimental reference. At present, the calculation steps of high-throughput screening are mainly composed of the following parts: select appropriate data from various databases; generate input data; select appropriate the algorithm performs performance calculation simulation; generates output files and performs data analysis; adds the processed data to the existing database for expansion, as shown in Figure 4.

![Flowchart of Integrated High Throughput Screening and Computational Analysis to Discover ST2 Inhibitors](image)

**Figure 4.** Flow chart of high-throughput screening

4. Conclusion

Machine learning is an emerging branch in the field of artificial intelligence. It has great potential in the field of materials science. Due to the flexibility, accuracy and generalization ability in dealing with...
complex system problems in actual situations, machine learning is worthy of researchers' full Pay attention. The application of machine learning is highly interdisciplinary. The application of machine learning in materials science is very broad, including metal materials, inorganic oxide materials, organic materials and various functional materials. Various related research It shows that machine learning can be used as a fast and accurate tool in the field of materials science. With the further development of theories and methods, machine learning will surely have more in-depth and extensive applications in the field of materials science.

Acknowledgments
This work was financially supported by Scientific research project of Bengbu University (2020ZR10, 2020ZR09).

References
[1] Liu, Y., Zhao, T., Ju, W., & Shi, S. Materials discovery and design using machine learning. Journal of Materiomics, 3 (3) (2017) 159 - 177.
[2] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. Machine learning in materials informatics: recent applications and prospects. npj Computational Materials, 3 (1) (2017) 1 - 13.
[3] Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials, 2 (1) (2016) 1 - 7.
[4] Sanchez-Lengeling, B., & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361 (6400) (2018) 360 - 365.
[5] Oliynyk, A. O., & Mar, A. Discovery of intermetallic compounds from traditional to machine-learning approaches. Accounts of chemical research, 51 (1) (2018) 59 - 68.
[6] Cao, B., Adutwum, L. A., Oliynyk, A. O., Luber, E. J., Olsen, B. C., Mar, A., & Buriak, J. M. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. ACS nano, 12 (8) (2018) 7434 - 7444.
[7] Kim, E., Huang, K., Saunders, A., McCallum, A., Cedar, G., & Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. Chemistry of Materials, 29 (21) (2017) 9436 - 9444.