Yuan *et al. J Mater Inf* 2024;4:14
DOI: 10.20517/jmi.2024.15

Journal of Materials Informatics

Review Article

Open Access

Check for updates

Applications of machine learning method in high-performance materials design: a review

Junhao Yuan[1,2], Zhen Li[3], Yujia Yang[1], Anyi Yin[1,*], Wenjie Li[4], Dan Sun[4], Qing Wang[2,*]

[1]Institute of Materials, China Academy of Engineering Physics, Jiangyou 621908, Sichuan, China.
[2]School of Materials Science and Engineering, Dalian University of Technology, Dalian 116024, Liaoning, China.
[3]School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, Liaoning, China.
[4]Science and Technology on Reactor System Design Technology Laboratory, Nuclear Power Institute of China, Chengdu 610213, Sichuan, China.

[*]**Correspondence to:** Prof. Qing Wang, School of Materials Science and Engineering, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian 116024, Liaoning, China. E-mail: wangq@dlut.edu.cn; Dr. Anyi Yin, Institute of Materials, China Academy of Engineering Physics, Jiangyou 621908, Sichuan, China. E-mail: anyiyin@126.com

How to cite this article: Yuan J, Li Z, Yang Y, Yin A, Li W, Sun D, Wang Q. Applications of machine learning method in high-performance materials design: a review. *J Mater Inf* 2024;4:14. https://dx.doi.org/10.20517/jmi.2024.15

Received: 6 Jun 2024 **First Decision:** 9 Jul 2024 **Revised:** 8 Sep 2024 **Accepted:** 20 Sep 2024 **Published:** 27 Sep 2024

Academic Editor: Hao Li **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

Abstract

As a generalized method of mathematical statistics, machine learning (ML) is playing an increasingly significant role in the realm of materials design. More sophisticated methodologies for in-depth understandings and wide applications have been developed from initially simple data relation mappings. The present work first summarizes the basic technical issues of ML and then systematically reviews the main implementation strategies for ML methods in accelerating materials research and development process in recent years, encompassing three primary aspects. Firstly, it is necessary to establish the relationship between the key characteristic parameters and properties in any given materials system for a better prediction and exploration of new materials. Then, the computational algorithms in materials science need to be optimized to replace complex calculations with model-predicted data. Finally, the ML methods are applied to summarize the one-dimensional property data and two-dimensional microstructural images of materials to establish standardized analysis methods. During this process, the domain knowledge in a specific system is of great significance to improving the prediction accuracy and efficiency of ML methods, whether pre-processing experimental or computational databases. The powerful capability of ML methods to handle high-dimensional data will enable researchers to make more effective decisions in materials design. In the future, the relationship between the microstructure and mechanical properties, which is necessary to establish a more effective search engine for alloys with targeted mechanical properties, will be the focus of ML mechanical properties of alloy materials.

Keywords: Machine learning, materials design, domain knowledge, applications of machine learning

INTRODUCTION

Machine learning (ML) is a scientific discipline, in which statistical models are constructed based on available data and applied to analyze the data and predict unknown data[1]. The most crucial core of ML is to develop various algorithms for data analysis and processing, which have been continuously refined since 1980[2-5]. The object of ML models is the "data" in the field of materials science, including the chemical compositions, characteristic parameters for representing microstructure and processing, metallographic images, *etc*.[6-9]. When the amount of data is abundant enough, they will exhibit a certain statistical regularity and render the ML model with a high prediction accuracy. Despite the availability of extensive databases with characteristic parameters in materials science, it is still a challenge to construct an explicit function that accurately reflects the specific physical meanings, resulting in a declined prediction accuracy. Recently, ML has emerged as a critical tool for managing vast datasets and optimizing parameters within the realm of Materials Genome Engineering (MGE), representing the "fourth paradigm" of data-driven strategies in the discovery and development of high-performance novel materials[10]. Materials design has undergone the evolution from empirical science (1st paradigm) to model-based theoretical science (2nd paradigm) and then to computational science (3rd paradigm)[11]. However, even with the assistance of computer technology, traditional methods for materials design still require researchers to make judgments on data interpretation and application. As the data dimensions increase, it becomes more challenging for researchers to manually identify the mapping relationship of key targets from multi-dimensional data, such as chemical composition, processing, and microstructure. Actually, the ML method is particularly well-suited for addressing these types of problems, which is known as the data-driven science (4th paradigm)[12].

In 2016, Raccuglia *et al*. established a reaction model for the crystallization process of inorganic-organic hybrid materials using ML based on failed or unsuccessful experimental data, in which the accuracy for predicting the formation conditions of new compounds reaches up to 89%[13]. This suggests that the advent of ML-assisted materials discovery has great potential to radically transform traditional approaches. Subsequently, the ML methods have attracted more attention and have been widely applied to both the development of new materials and the perdition of their macroscopic properties. Medasani *et al*. integrated the high-throughput first-principles calculations with the ML technique to predict the point defect behavior of binary intermetallic compounds with a B2 crystalline structure[14]. Takahashi *et al*. obtained a big database of interatomic potentials calculated by the density functional theory (DFT), and then constructed the atomic potential functions by ML with a linear regression method, which was well applied to the molecular dynamics (MD) simulation of Ti element[15]. So far, the ML has covered a broad range of material systems, including perovskite oxides, inorganic composites, superhard nitride and carbide ceramics, shape memory alloys, multi-principal-element high-entropy alloys (HEAs), Ni-base and Co-base superalloys, ultra-high-strength maraging stainless steels, high-strength and conductive copper alloys, *etc*.[6,16-26]. For example, the prediction of physical properties of perovskite materials, including band gap, stability, and electronic transport properties, has played a guiding role in the design of new materials[27]. The further development of ML will improve the interpretability of models, aid experimental research, and solve multi-scale problems. It will be of great significance to integrate ML with materials science and engineering education[28].

From the viewpoint of materials design, it is known that most high-performance materials were developed with the guidance of empirical methods and theories. For instance, the phase computation methods based

on *d*-electron theory have been playing an important role in the composition design to stabilize the specific face-centered-cubic (FCC)-γ/γ' coherent microstructure and inhibit the precipitation of brittle topological-close-packed (TCP) phases during the development of Ni-base superalloys[29]. The equivalent method, such as the Mo_{eq}, Cr_{eq}, or Ni_{eq}, was often used to characterize the structural stability of the parent phase for the prediction of the phase transformation in body-centered-cubic (BCC)-based Ti/Zr alloys and various stainless steels[30,31]. In practice, it significantly reduces the cost of massive trial-and-error experiments[32,33]. With the proliferation of computer technology in the field of materials science, the calculation of phase diagram (CALPHAD) method based on thermodynamic and kinetic principles has become prevalent. It can establish the correlation among chemical composition, phase constitution, and macroscopic properties (including mechanical, thermal, and electrochemical properties) in complex systems, resulting in the composition optimization and performance enhancement[34]. Thus, the screening of materials has evolved from the blind global search to the optimization of key parameters in multi-component systems. However, it is more difficult to find the mapping correlation among the key target parameters in the multi-dimensional data, such as composition, processing, microstructure, *etc*.[35]. Fascinatingly, the ML method is anticipated to deal with such challenges and has exhibited excellent adaptability.

As a mathematical method, ML can support the development of materials science in diverse manners due to its environmental self-adaptability[36]. Its primary task is to construct the mapping relationship between the input and the output, including the symbolic regressions for physical equations, the direct predictions of material properties for a given composition or process, *etc*.[6,37-39]. Intrinsically, it is a more generalized nonlinear-fitting method that can handle higher-dimensional data to enhance the analysis and prediction. However, it is emphasized that the sample database in any given system is so small that it can strongly affect the prediction accuracy of data-driven ML models[38,40-43]. The existing researches indicated that the determination of characteristic parameters to correlate the input and the output is crucial for improving the prediction accuracy of ML models containing a small database[44]. Meanwhile, the domain knowledge is typically employed to pre-process the data prior to the ML, in order to make full use of the limited but reliable experimental and computational data[44,45]. Therefore, the present work will summarize the applications of ML methods in materials science from three main aspects, being the development of high-performance materials using ML-assisted design models, the acceleration of calculations and simulations through ML, and the materials informatics guided by ML. Additionally, the role of domain knowledge in data pre-processing will be generalized to illustrate how to deal with the data information in the database, whether from experiments or calculations, for a better improvement of prediction accuracy.

This work provides a comprehensive review of the primary applications of ML methods in material design in recent years, focusing on the design of alloys and solid solution compounds. It begins by summarizing the technical issues that need to be considered when applying ML as a tool in materials science. Subsequently, it introduces the main forms of application of ML methods, including direct assistance in material design, acceleration of computational simulation work, and the construction of materials informatics. From a scientific perspective, researchers prefer models that are not just "black box" for data processing, but rather models that are interpretable or incorporate domain knowledge inherent to materials. This will play a significant role in the generalization of ML models.

TECHNICAL ISSUES OF MACHINE LEARNING FOR APPLICATION IN MATERIALS SCIENCE

Although the ML methods have important implications for materials design, there are still many issues that researchers need to pay attention to. The database construction, model selection, training strategies, and model evaluation are necessary technical elements of ML efforts. Each element will be discussed in this section.

Currently, the application of ML in the field of materials heavily relies on the construction of databases, such as Materials Project[46] and JARVIS[47], which can provide basic data of more than 100,000 different materials. With high-throughput computing, future ML tasks will be able to easily access over 10,000 pieces of data as a computational database for specific needs. However, it is important to emphasize that the experimental data is valuable due to the highest quality and information of experimental results. Thus, it can be used as the verification of ML results based on a computational database. Also, the data used in ML are limited by the generality of the data given by different research works. In practice, the construction of databases for specific ML tasks is closely related to the needs and capabilities of researchers. Therefore, a large amount of high-quality experimental data and integrated existing data will lead to better-performing ML methods.

Besides the ML database, there exist a series of discussions on the selection of ML models for better utilization. In the early days, software packages, such as Scikit-learn[48] and XGboost[49], were commonly used for ML tasks. In recent years, several integrated ML frameworks developed by large companies have emerged, including TensorFlow[50], PyTorch[51], AutoGluon-Tabular[52], *etc*., which have been significantly optimized in terms of usability and versatility. In the field of materials science, several neural network-based models, such as Crystal Graph Convolutional Neural Networks (CGCNN)[53], MatErials Graph Network (MEGNet)[54], and DeePMD-kit[55], have made breakthroughs in their respective targeted areas, whether for important parameters such as electrical conductivity or for constructing ML potential functions. With the increasing popularity of ML in materials science, researchers can choose a well-established targeted toolkit or a universal ML algorithm, based on which some modifications could be made according to their specific requirements.

Notably, it is not just about selecting algorithms and toolkits during ML. Researchers should also evaluate the ML methods by considering their ability to interpret data. Wang *et al*. systematically discussed the ML methodologies ranging from supervised learning (SL) to transfer learning (TL) and unsupervised learning (UL)[56] and pointed out that both SL and TL require the construction of a vast foundational database for further design. The difference between them is that the TL can learn from existing data and then transfer model parameters or material features to the target domain/tasks, which makes the cost less than building a new database. Furthermore, the UL can well handle the complex correlations among data, which is beneficial for understanding the intrinsic structure-property relationships of diverse materials. A reasonable UL can significantly reduce the workload required to obtain material properties, shortening the design cycle and achieving a higher accuracy for more complex systems. Therefore, both the key features transferred by TL and the feature recognition constructed by UL are expected to provide interpretable mechanisms for data. To achieve this goal, the generalization ability and accuracy of ML models can be improved by embedding domain knowledge appropriately.

Another issue encountered during ML is how to define the accuracy of the model. This involves two aspects: how to train and validate the model and whether the selection of loss functions is reasonable. The training and validation of ML models are crucial to ensure their generalization ability. Common strategies involve five aspects. The first one is data pre-processing, which includes cleaning data, handling missing values, feature encoding, and normalization. Secondly, the data set needs to be divided into training, validation, and testing subsets for the evaluation of model's performance on unseen data. Thirdly, the K-fold method is applied to cross-validate the ML model for improving the stability and reliability of model evaluation. Then, a regularization is performed to reduce the model complexity and avoid overfitting. The

final strategy is the hyperparameter tuning, where several methods such as grid search, random search, or Bayesian optimization can be used to achieve the optimal model parameters. To obtain an accurate ML model, these strategies can be comprehensively applied in the model-building process.

In addition, loss functions are often employed to characterize the ML accuracy, but the applicability of these loss functions has not been well-assessed in previous studies. Naser and Alavi analyzed 78 common loss functions to illustrate their applicability in engineering and scientific problems[57], and they also advocated the use of Performance Fitness and Error Metrics (PFEMs) as the criterion for selecting loss functions. Unfortunately, many published studies on ML applications in engineering do not include multi-criteria or additional validation stages; instead, they rely solely on traditional performance metrics, such as R or R^2 of derived models. Also, a set of PFEMs does not completely eliminate some common problems, in which the over-fitting and bias are the most obvious. Thus, both the learning strategies and loss functions play a decisive role in whether the ML method is reasonable.

ML-ASSISTED DESIGN MODELS FOR DEVELOPING HIGH-PERFORMANCE MATERIALS

The most essential issue of ML-assisted materials design is to determine the key system parameters, which can not only affect the macroscopic properties but also characterize the composition, microstructure, and processing. With the guidance of these key parameters to constrain the input and the output, the prediction accuracy of the ML model can be improved significantly during the development of novel high-performance materials in any given system. Generally, the construction of such ML-assisted design models can be outlined into six steps: establishing the database, determining the key parameters, training and testing the ML models, designing new materials by the ML, verifying the properties of designed materials by experiments, and, finally, integrating the obtained data back into the database to initiate a new iteration of ML. The framework of the whole ML is presented in Figure 1. Through several iterations, the mapping relationship between the input and the output will progressively become more precise, which can undoubtedly accelerate the research and development of novel materials with a higher efficiency. In the following, we will provide five representative examples to illustrate the applications of ML method in the development of novel materials.

HEAs with prominent properties

Unlike traditional alloys that are typically dominated by one or two base elements, the HEAs are composed of multiple principal components with an equimolar or non-equimolar mixing, which can provide a tremendous composition space for exploring high-performance alloys[58]. It is due to the vast composition space that can render HEAs with prominent mechanical and functional properties simultaneously[59]. However, it is a great challenge to design new HEAs because of the limited quantity of data available in existing databases. In this case, Wen *et al*. used several key parameters, including the itinerant electrons *e/a*, modulus mismatch, work function, *etc*., to verify the phase constitution of HEAs in the Al-Co-Cr-Cu-Fe-Ni system, which can well bridge the correlation between the input (chemical composition) and the output (microhardness HV) of ML models with a database containing 155 samples[20]. Thus, the alloy design strategy combining the ML model with the experimental design rules was formulated to search for new HEAs with a much larger HV. It was found that the HV values of 35 alloys in 42 newly-designed compositions are higher than the maximum, and the hardness of 17 alloys is enhanced by more than 10% compared to the maximum in the training dataset. This indicates that the strategy incorporating both the composition and key parameters performs better than the one that solely relies on composition.

Furthermore, Liu *et al*. have developed a ML-guided high-throughput experimental approach to expedite the development of non-equimolar superhard Co-Cr-Ti-Mo-W HEAs, which offers an effective strategy

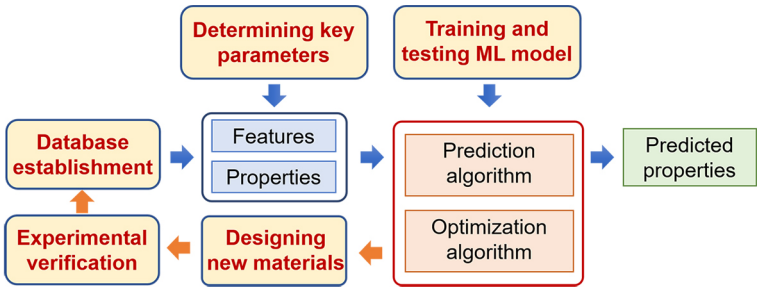

Figure 1. The framework of materials design process by machine learning.

that has the potential to increase overall efficiency by a hundred-fold and reduce costs significantly, compared with conventional methods[60]. Specifically, the final ML model, trained using 138 experimental data, can predict the alloy hardness with the mean relative errors of 5.3%, 6.3% and 15.4% in the high (HV > 800), medium (HV = 600-800), and low (HV < 600) hardness ranges, respectively. Among them, 14 superhard HEAs with HV > 900 were discovered by the ML-guided high-throughput experiments. Moreover, multiple ML models were used to predict the hardness of 3,876 hypothetical alloys covering the entire composition range. This analysis revealed the systematic composition effects based on the comprehensive correlations between composition-hardness and descriptor-hardness, where the descriptors include several crucial parameters, such as the valence electron concentration (VEC), melting temperature (T_m), enthalpy of mixing (ΔH), entropy of mixing (ΔS), and atomic size difference (δ).

Besides mechanical hardness, the thermodynamic properties, such as the thermal expansion coefficient (TEC), are also important for the application of HEAs. Rao *et al.* proposed an active learning strategy to accelerate the design of high-entropy Invar alloys in a practically infinite composition space, even with limited data available[61]. This approach operated as a closed-loop system, seamlessly integrating the ML with the density-functional theory, thermodynamic calculations, and experimental validation, which can achieve an accurate prediction of properties across a wide compositional space. After processing and characterizing 17 new alloys out of millions of possible compositions, two high-entropy Invar alloys were then identified with an extremely low TEC of ~ 2×10^{-6} per degree kelvin at 300 K.

High-strength and conductive Cu alloys

High-performance Cu alloys play a fundamental role in the integrated circuit and railroad industries, which need to possess both high strength and high electrical conductivity (EC) for meeting the application requirements. However, a higher strength always corresponds to a lower EC in most existing Cu alloys. Wang *et al.* proposed a property-oriented design strategy for high-performance Cu alloys via the ML method, which involves three crucial features, including ML modeling, compositional design, and property prediction[62]. This constructed ML model exhibits better efficiency in the achievement of a rapid composition design of Cu alloys with a targeted ultimate tensile strength of UTS = 600~950 MPa and an electrical conductivity of EC = 50% international annealed copper standard (IACS).

In order to further enhance the efficiency of composition design, all alloy parameters were subjected to correlation screening, recursive elimination, and exhaustive screening during the ML process. Subsequently, the composition was iteratively designed through Bayesian optimization[63]. Thus, five kinds of key parameters affecting the microhardness HV and six kinds of key parameters affecting the EC were obtained by screening out alloy parameters to establish the "HV - key parameters" model and "EC - key parameters" model, respectively, where the accuracy of these two models exceeds 90%. Then, novel Cu alloys were effectively designed using the Bayesian optimization and iterative optimization experiments. Among them,

the designed Cu-1.3Ni-1.4Co-0.56Si-0.03Mg (wt.%) alloy has prominent properties with UTS = 858 MPa and EC = 47.6% IACS, being superior to those reported results in precipitation-strengthened Cu alloys. Therefore, this approach breaks through the dilemma between the strength and electrical conductivity.

Multi-component β-Ti alloys with low Young's modulus

BCC β-Ti alloys with low Young's modulus (*E*) were always achieved by co-adding BCC-stabilized elements (Mo, Nb, Ta) and low-*E* elements (Zr, Sn)[31,64]. When the BCC structural stability is not enough or too high, the *E* values of alloys will increase. Also, the precipitation of some metastable phases (α', ω, *etc.*) caused by an inappropriate matching among alloying elements enhances the *E* of alloys. The Mo equivalence (Mo_{eq}) was often used to characterize the BCC structural stability, and the low *E* could be obtained at the lower limit of Mo_{eq}. In our previous work, we proposed a cluster-plus-glue-atom model to describe the chemical short-range orders (CSROs) induced by the solute atoms in solid solution structure, from which a cluster composition formula could be abstracted[65]. In particular, the cluster formula in Ti-Zr-Mo-Sn-Nb-Ta system was expressed with $[(Mo,Sn)\text{-}(Ti,Zr)_{14}](Nb,Ta)_{1\sim3}$ to determine the added amount of each alloying element[66,67], which can be taken as the composition constrained parameter for ML. Thus, we implemented the Mo_{eq} and cluster formula into the ML model to design and predict novel multi-component low-*E* β-Ti alloys [Figure 2]. Both auxiliary gradient-boosting regression tree and genetic algorithm methods were adopted to deal with the optimization problem in the ML model[68]. This cluster-formula-embedded ML model can not only predict alloy properties in the forward design, but also design and optimize alloy compositions with desired properties efficiently and accurately. By setting different objective functions, only several (3~5) new β-Ti alloys with either the lowest *E* (*E* = 48 GPa) or a specific *E* (*E* = 55 and 60 GPa) were predicted by ML and then validated by a series of experiments, from which it could be found that the experimental *E* can be well consistent with the predicted one. Here, it is necessary to emphasize that if the cluster formula was not embedded in the reserve design of ML, 85 alloy compositions could be predicted by ML for a specific *E* = 55 GPa, which inevitably intensifies the difficulty in experimental verification. So, the cluster-formula-embedded ML model can make the prediction and optimization of composition and property more accurate, effective, and controllable, since the composition constraint was implemented to reduce the composition variants.

Novel high-strength Co-base superalloys with γ/γ' coherent microstructure

For mature Co-based superalloy systems, the thermodynamic databases can serve as a basis for designing of new alloys. Due to the huge amount of data in the database, the accuracy of ML models can be ensured from the source of the database. This vast data can support ML methods to optimize multiple targets simultaneously. Based on this, Liu *et al.* proposed a materials design strategy to optimize multiple targeted properties of multi-component Co-base superalloys via ML[6]. The microstructural stability of coherent γ and γ' phases, the solvus temperature and volume fraction of γ' phase, alloy density, processing window, freezing range, and oxidation resistance were optimized at the same time. Finally, a series of novel Co-base superalloys were successfully designed and experimentally synthesized from more than 210,000 candidates. The best performer, Co-36Ni-12Al-2Ti-4Ta-1W-2Cr (at.%), possesses the greatest γ/γ' microstructural stability without any precipitation of deleterious phases, the highest γ' solvus temperature of 1,266.5 °C, a higher volume fraction (74.5%) of γ' nanoparticles, a relatively-lower density of 8.68 $g{\cdot}cm^{-3}$, as well as excellent oxidation resistance at 1273 K due to the formation of protective alumina scales.

Inorganic compound materials

Pb-free $BaTiO_3$ (BTO)-based piezoelectrics with a large electrostrain have great potential in the applications of micromotors and prosthetic devices[16,69]. Generally, a large electrostrain was achieved by chemical substitution (i.e., dopants) with Ca^{2+} and Sr^{2+} cations for Ba^{2+}, and with Zr^{4+} and Sn^{4+} for Ti^{4+} in BTO. This family is always expressed with the formula of $(Ba_{1.0-x-y}Ca_xSr_y)(Ti_{1.0-u-v}Zr_uSn_v)O_3$, showing a vast search space

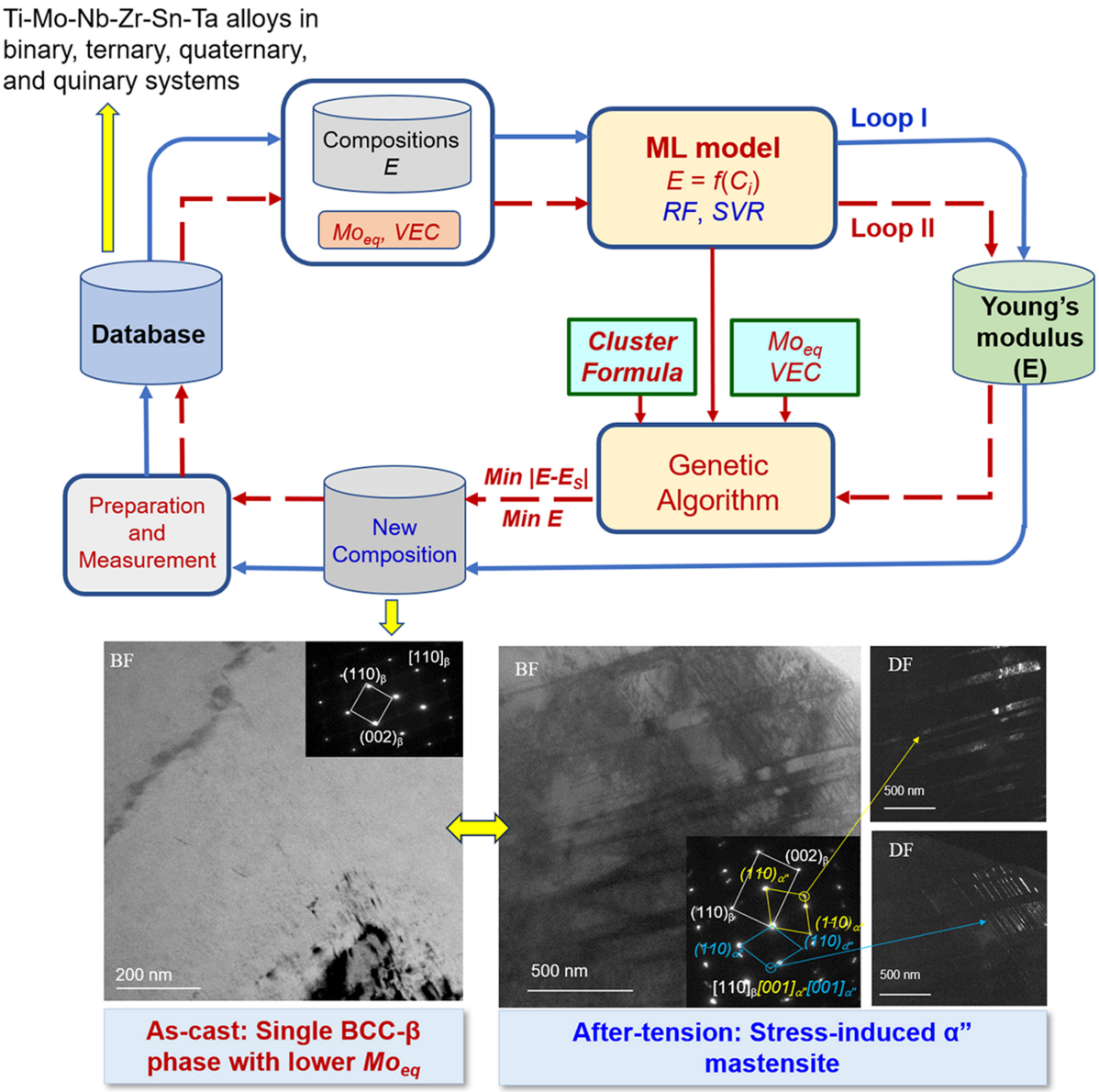

Figure 2. Cluster-formula-embedded machine learning for the design of low-*E* β-Ti alloys. The images are adapted from ref[68] with permission.

that cannot be explored by the trial and error or the intuition alone. Indeed, the key challenge in guiding experiments toward materials with desired properties lies in the effective navigation of the extensive search space encompassing the chemistry and structure of permissible compounds. Yuan *et al.* coupled the ML with optimization methods to accelerate the discovery of novel BTO-based piezoelectrics[16]. By experimentally comparing several design strategies, it is found that the active learning approach can balance the trade-off between the exploration (using uncertainties) and exploitation (using model predictions alone) and then obtain the optimal criterion, resulting in the synthesis of $(Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$ with the largest electrostrain of 0.23% in the BTO family.

Similarly, for other specific systems, such as uranium dioxide (UO_2) composite fuels, the ML method can be applied to construct the correlations among computational data to accelerate design efficiency since the amount of experimental data is limited. The lower thermal conductivity (TC) of UO_2 needs to be improved by injecting a second phase with a high TC into the matrix. Yan *et al*. utilized the finite element method (FEM) to generate massive simulated measurements and then proposed a novel algorithmic method to learn automatically from gathered simulation results[70]. Through the neural network, a set of key features associated with the effective thermal conductivity (ETC) in 2D-FEM were found to achieve both the forward and reverse predictions. Then, the correlation in 2D space was extended to 3D space for realizing the high-precision prediction of calculated results of 3D-FEM. With the guidance of the model, not only the ETC of a composite fuel can be predicted accurately according to its given structural characteristics, but also the structural characteristics of a composite fuel could be inferred from its expected ETC, in which the relative error of forward prediction and inverse design is less than 5%.

ML-ACCELERATED CALCULATIONS AND SIMULATIONS

Another important application of ML is to simplify and optimize the computational models for fundamental physics. For example, in the process of obtaining the numerical solution of the function, every minor adjustment of parameter needs to be re-iterated. Thus, it is often necessary to modify parameters several times during the conduction of a comprehensive complex calculation, which can be optimized with the ML models by fitting a series of the input and output. So, when the next calculation for the parameter adjustment is performed, the result could be predicted directly through the ML model instead of another complex calculation. If the number of calculations is sufficiently large, it will significantly save time in subsequent calculations after training a highly accurate model. In the first-principles (*Ab-initio*) simulation software such as the Cambridge Sequential Total Energy Package (CASTEP) and the Vienna Ab initio Simulation Package (VASP)[71,72], the on-the-fly method based on the ML potential function has been integrated. For the *Ab-initio* molecular dynamics (AIMD) calculations enabled by the on-the-fly method, the output data can be added into the ML model to fit the potential function. When the fitting result reaches a certain accuracy, the ML potential will replace the self-consistent calculation for subsequent calculations, in which the Bayesian error is used to monitor the potential function. If the error is too high, a self-consistent calculation step will be added for correction. Finally, an efficient and accurate potential function could be obtained after a series of iterations[73,74]. As an instance, the superiority of potential functions from ML force field was verified in the solid and liquid models of several systems, such as Al, Sn, Ge, Sn, MgO, *etc*.[73]. In a lot of calculations, the ML has replaced 99% of the first-principles calculations, resulting in an efficiency improvement of over 200 times. The average absolute errors [or the root mean square errors (RMSE)] for energy, force, and stress tensor are only 5.5 (6.2) meV·atom^{-1}, 0.07 (0.09) eV·Å^{-1}, and 0.18 (0.27) GPa, respectively.

In addition to the applications of AIMD, the potential functions can be further abstracted and used in large-scale MD calculations[75]. The MD calculations are often limited by potential functions and always need extensive efforts to construct a reliable potential function from the basic physical theorem, which is time-consuming and laborious. The ML models could be applied to fit the input and output of physical formulas for obtaining high-precision potential functions in any specific system. For instance, Zong *et al*. established a domain-knowledge-embedded database containing structures and properties according to the existing physical basis and the background of specific material systems[76]. Furthermore, the ML and the neural network hybrid method has been developed to construct a more accurate interatomic potential function by learning from the AIMD simulations and the domain-knowledge-embedded database. On this basis, the developed ML potential function could be used to simulate the microstructural evolution and predict physical properties of allotropic metals under the static high pressure and dynamic impact environments,

where the microscopic mechanisms can be well understood in combination with the domain knowledge. This ML-AIMD interatomic potential accurately captures the energetics and structural transition properties of zirconium, compared with the experimental results and the density functional data for phonons, elastic constants, and stacking fault energies. The maximum absolute error of ML-AIMD energy is less than 6.7 meV/atom, showing a high reliability. Furthermore, the MD simulations using this potential function successfully reproduce the phase transformation mechanism of zirconium and draw the pressure-temperature phase diagram of zirconium.

ML-BUILD MATERIALS INFORMATICS

Intrinsically, the ML models act as "black boxes" to simplify complex physical calculations and to construct associations among huge data. However, when it was used to explain the mapping relationship between data, the "black box" should be simplified as much as possible into a formulaic model[77]. Based on the empirical rules of existing formulas, the algorithm framework and parallel program have been developed based on data searching formulas. Firstly, the algorithm requires the user to specify the main variables in the given system, and write them as $y = f(x_1, x_2, ..., x_n)$ along with the dimensions of variables. Then, the program automatically performs dimensional analysis to build the feature space. After iterative filtering, the program carefully uses the well-designed scoring rules to balance the accuracy of the formula in describing the data and the simplicity of the formula itself, and finally outputs the best formulas as the alternative solutions. In addition to the traversal search method, the program also provides the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm and Genetic Algorithm (GA) as the alternatives in the vast space to find alternative solutions for the optimal. Wei *et al.* utilized this set of formulas to construct a flow and to establish highly interpretable formulas for describing the weight gain changes of FeCrAlCoNi HEAs exposed to temperature and humidity in the air, where the average error is $R^2 = 0.962$, indicating a high level of reliability[78].

Besides the construction of one-dimensional data relationship mapping, the ML methods represented by neural networks and deep learning models can also process and identify images. For the alloy design, the image information, such as the microstructure, is of great significance. As a bridge between composition/process and mechanical properties, the microstructure quantification is an important tool for understanding physical metallurgy mechanism. Shen *et al.* proposed a deep learning approach to train electron backscattered diffraction (EBSD) images by integrating advanced materials characterization techniques and artificial intelligence strategies[7]. The workflow is shown in Figure 3. In this model, the EBSD characterization was employed to obtain accurate phase classification labels from microstructures. The U-Net deep learning architecture for small sample data was applied to establish a high-dimensional mapping correlation between the scanning electron microscope (SEM) morphology and phase diagrams. Data sets of dual-phase (DP) steel and quenching and partitioning (Q&P) steel were constructed through systematic characterization experiments. The original image was segmented into several sub-images with a size of 128 × 128 pixels, where some data enhancement methods such as flipping were also taken to expand the data size. Finally, the datasets of these two kinds of steels contain 1,914 and 6,048 sub-images respectively, in which both the compression and expansion paths consist of four convolutional layers, respectively. Thus, the optimal model of the microstructural segmentation under each data set was obtained by adjusting the model parameters, such as the loss function and optimizer.

APPLICATIONS OF DOMAIN KNOWLEDGE DURING MACHINE LEARNING

With the widespread application of ML methods, it is gradually realized that the data available for any specific materials system is scarce and highly valuable. In order to construct a more reliable ML model, the knowledge background behind these rare data in the algorithms is of great importance for improving the

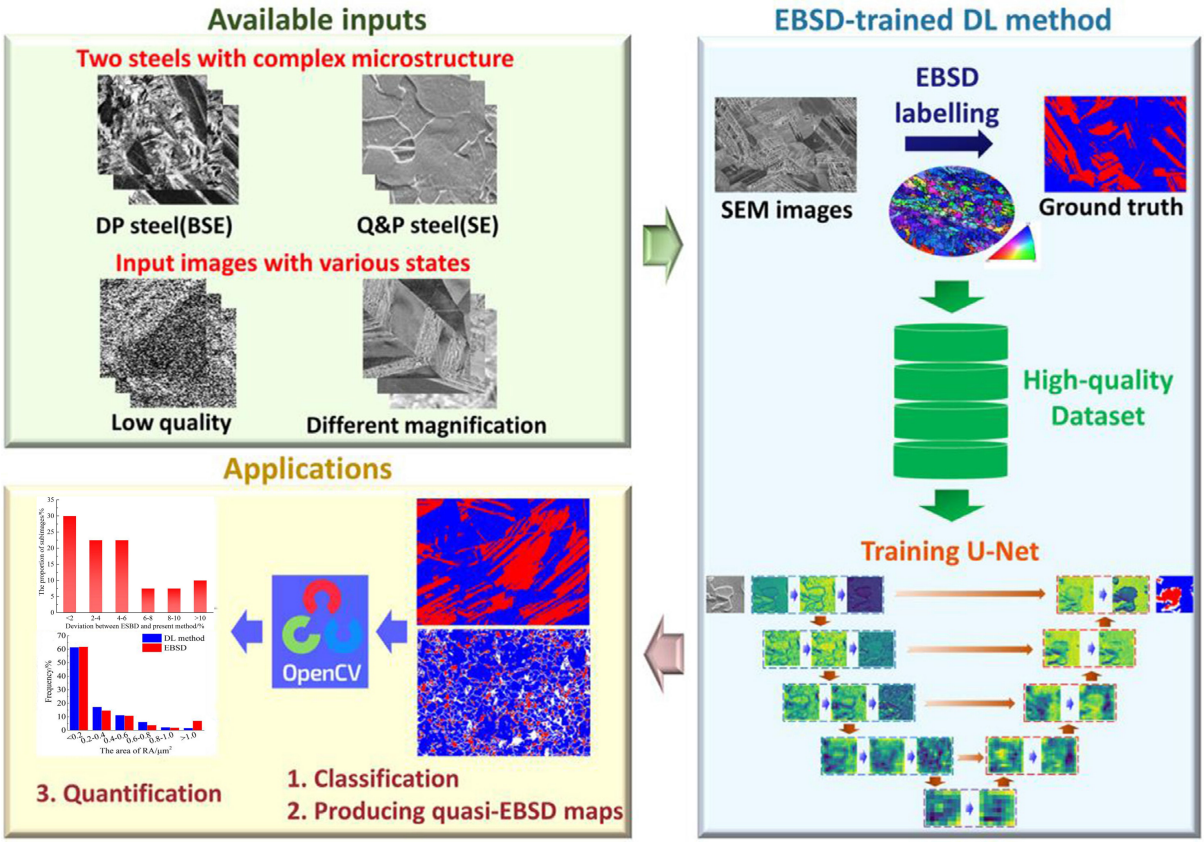

Figure 3. Schematic diagram of the EBSD-trained deep learning method and its application in dual-phase steel and quenching and partitioning steel. The images are quoted from ref[7] with permission. EBSD: Electron backscattered diffraction.

model accuracy. Besides the direct modeling of experimental data, a large number of computational data containing domain knowledge can be obtained using computational methods with the basic information input such as physical formulas. As one of the application ways of domain knowledge, the experts system has long been widely used in the evaluation and diagnosis field of civil engineering systems for evaluation and diagnosis[79]. Similarly, in materials science, domain knowledge that greatly enhances ML will be a great enhancement to ML algorithms in materials science. So, the introduction of domain knowledge is a significant approach to enhance the data quality for a better prediction of ML.

A modular process for embedding domain knowledge into ML

For initial data in different systems, the introduction of domain knowledge into ML models can be divided into several modules, including the database construction and pre-processing, ML model construction and hyper-parameter optimization, forward prediction, and reverse design [Figure 4]. As the key step for building an optimal ML model, the high-quality database should be ensured firstly. The database is usually sourced from two kinds of data, and one of them is the experimental data, which is difficult to obtain. In fact, there always existed a relatively-large error among these experimental data due to the variations in the preparation of alloy samples by different researchers, such as the processing and composition deviations. Domain knowledge can be employed to pre-process these data, including the classification, grading, weighting, *etc*. The other kind of data were obtained by targeted computational simulations, where the domain knowledge can reduce computational effort and ensure its rationality. Thus, the domain knowledge achieved from materials science can be well used to optimize the dataset.

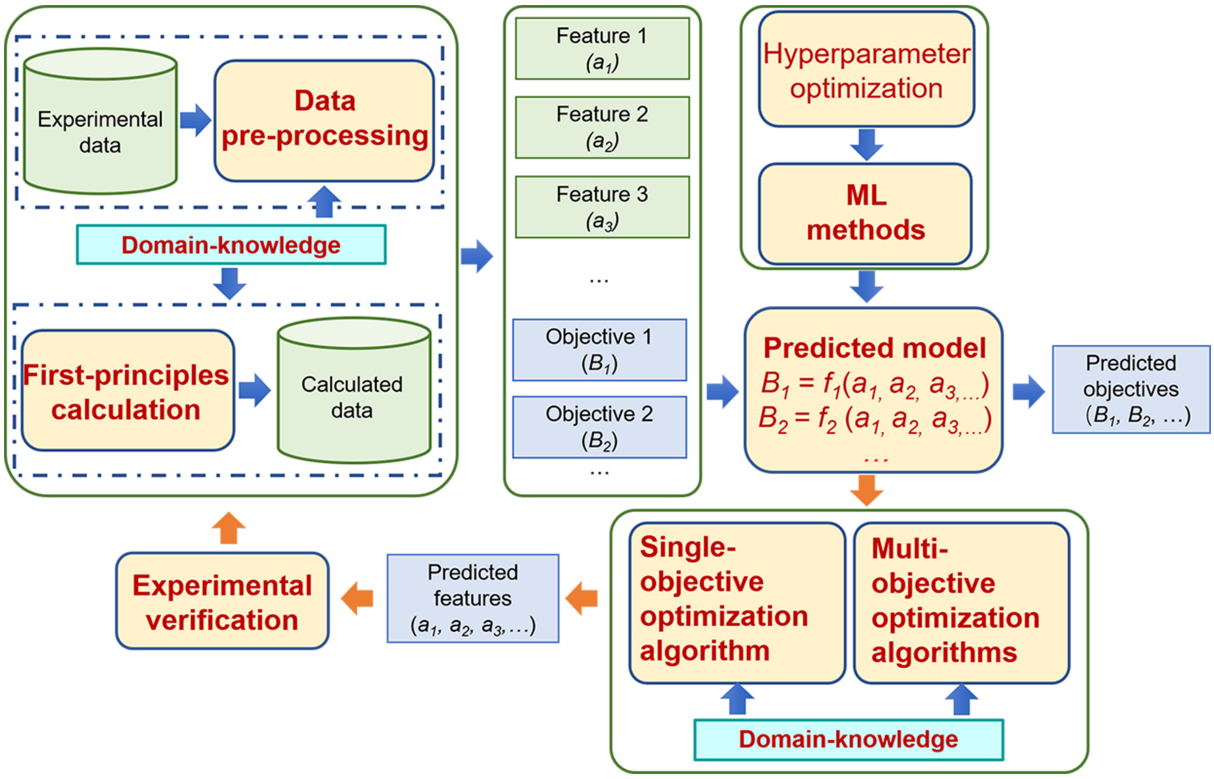

Figure 4. The modular process for domain-knowledge-embedded machine learning.

Here, the property-orientated alloy-design strategy, combining the ML and the characteristic parameters, was taken for an instance to search for the BCC γ-U alloys with prominent corrosion resistance in the U-Mo-Nb-Ti-Zr systems[80]. To enhance the accuracy of predictions, the cluster formula approach and Molybdenum equivalent (Mo_{eq}) were embedded as domain knowledge of the ML model. The cluster formula, which reflects the elemental interactions, acts as compositional constraints, while the Mo_{eq} signifies the structural stability of the BCC phase. Before ML, domain knowledge was leveraged to pre-process the data, including data screened and weighted, to enhance the model's predictive power. Armed with a screened and weighted dataset, Auxiliary Gradient Boosting Regression Tree (XGBR) methods were employed to establish an optimal correlation between alloy composition and corrosion-resistant lifetime (D) in boiling water at 343 °C. This approach outperformed both the Random Forests Regressor (RFR) and Support Vector Regression (SVR) methods. Subsequently, the cluster formula was integrated into the ML model for the reverse design, aiming to forecast new alloys tailored to achieve a desired D. The ML model successfully designed a multi-component alloy with a composition of U-7.17Mo-0.96Nb-0.31Ti-0.28Zr (wt.%), exhibiting an extended corrosion-resistant lifetime with a maximum D (190.4 days). In the absence of cluster formula constraints on alloy compositions, ML would yield 158 potential compositions when targeting $D \geq 182.0$, significantly complicating experimental validation. Thus, the cluster-formula-embedded ML method was proved as an efficient tool for predicting alloy compositions in multi-component systems.

When the ML model is limited by the quantity and quality of experimental data, the domain knowledge can also be embedded into the computational methods in advance to obtain a series of reliable data with intrinsic characteristics for ML, which will open a breakthrough for the design of novel materials with scarce experimental data. Zou *et al.* developed a ML design method for high-strength and ductile Ti alloys based on high-throughput first-principles calculations[81]. Among them, three procedures, the formula

description of strengthening mechanisms, the empirical design principle, and the calculation of multi-component solid solution model by the similar atomic environment (SAE) approach[82], were required to obtain the meaning functions from the knowledge-based modeling, which is essential to further optimize the design strategies recommended by ML. After data mining and ML, the Ti-7543 (Ti-7Mo-5Al-4Cr-3Nb-0.5Fe-0.2O wt.%) alloy was obtained, exhibiting a higher yield strength of 1239 MPa and good ductility with the elongation of 8.2 %. If these underlying requirements and principles in the knowledge base were not considered, there would exist a huge deviation in the predicted strength data, such as the contributions from the solid-solution strengthening and grain refinement hardening, which will conflict with these related strengthening mechanisms.

Another way to enhance the generalization capability of ML models is to employ interpretable ML models. In Liu and Sun's work, the Explainable Boosting Machine (EBM) was utilized to forecast the compressive strength of concrete materials and to elucidate the contribution of mix ratio factors for the compressive strength[83], where 1030 compressive strength values were used to construct the model. The Bayesian optimization algorithm was applied to iteratively construct the hyperparametric model and identify the optimal point in the space, which significantly reduced the time consumption in building the ML model. In terms of model prediction performance, the EBM algorithm exhibited excellent performance with the coefficient of determination (R^2) being 0.93, the *RMSE* being 4.33, and the mean absolute error (*MAE*) being 3.10, respectively. It allows a comprehensive interpretation of the contribution of individual features to the predicted results from both global and local perspectives, thereby further determining the influence of each mixing ratio on the compressive strength of the concrete. It should be noted that the interpretable model itself does not involve the application of domain knowledge. However, the various parameters provided by interpretable models will ultimately affect the performance and serve as an important reference for materials design, from which the intrinsic mechanisms in materials can be gleaned as new domain knowledge.

CONCLUSIONS AND PROSPECTS

ML methods have provided new approaches for materials researchers to resolve present challenges. In practical applications, these methods can be used to construct mapping correlations, whether to formulate physical equations or directly to predict materials properties for any given compositions or processes. For materials design, the ML methods can identify the key parameters that determine performance and achieve design goals. In computational materials science, the ML methods can accelerate the construction of important calculation foundations, such as potential functions, thereby improving the speed of AIMD methods by one or two orders of magnitude. When describing a specific phenomenon, the ML can construct accurate and precise mathematical formulas. Moreover, deep learning methods can also be utilized to recognize the image information for accelerating the analysis processes. In situations where the data are scarce but valuable, the introduction of reliable domain knowledge into the ML can allow for a certain degree of independence from relying solely on big data. This approach will allow the materials design to be guided only by a few key parameters, thereby enabling a greater focus on improving precision and efficiency rather than relying on a large amount of data.

In the future work, the application of ML in material systems will involve two directions. First, from a technical point of view, the application of ML will become more standardized and regulated, including the standardization of database construction patterns and data formats, as well as the standardization of model selection and training methods. The standardization of databases will enable ML to have more and more data available in the future, thereby promoting the "data-driven" approach from the technical level. The standardization of model selection and training will automate the entire process of ML, starting from the database to efficiently obtain the most suitable model. Further, for the material systems, through the

accumulation of critical data, the ML methods have established an algorithm-based understanding of materials science. The powerful capability of these methods to handle high-dimensional data will enable researchers to make more effective decisions in materials design. The ML has reduced the enormous search space for functional materials and has facilitated the endless quest for improving novel materials. However, compared with the physical and chemical properties, the mechanical properties of alloy materials involve more factors, such as composition and microstructure. It needs to be emphasized that the microstructure of an alloy is largely determined by a series of processing factors. Thus, the relationship between the microstructure and mechanical properties remains unclear, and establishing this connection is crucial for developing a more effective search engine for alloys with targeted mechanical properties. This will be the focus of future ML applications in predicting mechanical properties of alloy materials[84].

Recently, "MatGPT" of materials informatics has been proposed, outlining technical roadmaps for data, descriptors, generative models, pre-trained models, directed design models, collaborative training, and experimental robots[85]. It aims to bring materials design work closer to the Large Models, thus achieving the "ChatGPT" in the field of materials. With the development of "MatGPT", materials researchers will finally be liberated from tedious, repetitive work and have more energy to conduct more creative research.

DECLARATIONS

Authors' contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: Yuan J, Wang Q, Li Z

Performed data acquisition and provided administrative, technical, and material support: Yin A, Yang Y, Li W, Sun D

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China (52171152 and 12205286).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Zhou ZH. Machine learning. Springier Nature; 2021. DOI
2. Mandal S, Sivaprasad PV, Venugopal S, Murthy KPN. Artificial neural network modeling to evaluate and predict the deformation behavior of stainless steel type AISI 304L during hot torsion. *Appl Soft Comput* 2009;9:237-44. DOI
3. Guo Z, Sha W. Modelling the correlation between processing parameters and properties of maraging steels using artificial neural network. *Comput Mater Sci* 2004;29:12-28. DOI
4. Gavard L, Bhadeshia HKDH, MacKay DJC, Suzuki S. Bayesian neural network model for austenite formation in steels. *Mater Sci*

Technol 1996;12:453-63. DOI
5. Bailer-jones C, Bhadeshia H, Mackay D. Gaussian process modelling of austenite formation in steel. *Mater Sci Technol* 1999;15:287-94. DOI
6. Liu Y, Wu J, Wang Z, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Mater* 2020;195:454-67. DOI
7. Shen C, Wang C, Huang M, Xu N, van der Zwaag S, Xu W. A generic high-throughput microstructure classification and quantification method for regular SEM images of complex steel microstructures combining EBSD labeling and deep learning. *J Mater Sci Technol* 2021;93:191-204. DOI
8. Zhang Z, Wen G, Chen S. Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding. *J Manuf Process* 2019;45:208-16. DOI
9. Ma J, Luo D, Liao X, Zhang Z, Huang Y, Lu J. Tool wear mechanism and prediction in milling TC18 titanium alloy using deep learning. *Measurement* 2021;173:108554. DOI
10. Su Y, Fu H, Bai Y, Jiang X, Xie J. Progress in materials genome engineering in China. *Acta Metall Sin* 2020;56:1313-23. (in Chinese). DOI
11. Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. *Adv Sci* 2019;6:1900808. DOI PubMed PMC
12. Agrawal A, Choudhary A. Perspective: materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater* 2016;4:053208. DOI
13. Raccuglia P, Elbert KC, Adler PDF, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* 2016;533:73-6. DOI
14. Medasani B, Gamst A, Ding H, et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *npj Comput Mater* 2016;2:1. DOI
15. Takahashi A, Seko A, Tanaka I. Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: application to elemental titanium. *Phys Rev Materials* 2017;1:063801. DOI
16. Yuan R, Liu Z, Balachandran PV, et al. Accelerated discovery of large electrostrains in $BaTiO_3$-based piezoelectrics using active learning. *Adv Mater* 2018;30:1702884. DOI PubMed
17. Wang C, Shen C, Cui Q, Zhang C, Xu W. Tensile property prediction by feature engineering guided machine learning in reduced activation ferritic/martensitic steels. *J Nucl Mater* 2020;529:151823. DOI
18. Niu B, Wang Z, Wang Q, et al. Dual-phase synergetic precipitation in Nb/Ta/Zr co-modified Fe-Cr-Al-Mo alloy. *Intermetallics* 2020;124:106848. DOI
19. Domínguez LA, Goodall R, Todd I. Prediction and validation of quaternary high entropy alloys using statistical approaches. *Mater Sci Technol* 2015;31:1201-6. DOI
20. Wen C, Zhang Y, Wang C, et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater* 2019;170:109-17. DOI
21. Chang YJ, Jui CY, Lee WJ, Yeh AC. Prediction of the composition and hardness of high-entropy alloys by machine learning. *JOM* 2019;71:3433-42. DOI
22. He L, Wang Z, Akebono H, Sugeta A. Machine learning-based predictions of fatigue life and fatigue limit for steels. *J Mater Sci Technol* 2021;90:9-19. DOI
23. Liu P, Huang H, Antonov S, et al. Machine learning assisted design of γ'-strengthened Co-base superalloys with multi-performance optimization. *npj Comput Mater* 2020;6:334. DOI
24. Zhang J, Xu B, Xiong Y, et al. Design high-entropy carbide ceramics from machine learning. *npj Comput Mater* 2022;8:678. DOI
25. Qiao L, Zhu J, Wan Y, Cui C, Zhang G. Finite element-based machine learning approach for optimization of process parameters to produce silicon carbide ceramic complex parts. *Ceram Int* 2022;48:17400-11. DOI
26. Xue D, Xue D, Yuan R, et al. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Mater* 2017;125:532-41. DOI
27. Xia K, Gao H, Liu C, et al. A novel superhard tungsten nitride predicted by machine-learning accelerated crystal structure search. *Sci Bull* 2018;63:817-24. DOI PubMed
28. Yu J, Xi S, Pan S, et al. Machine learning-guided design and development of metallic structural materials. *J Mater Inf* 2021;1:9. DOI
29. Matsugi K, Murata Y, Morinaga M, Yukawa N. An electronic approach to alloy design and its application to Ni-based single-crystal superalloys. *Mater Sci Eng A* 1993;172:101-10. DOI
30. Mehjabeen A, Xu W, Qiu D, Qian M. Redefining the β-phase stability in Ti-Nb-Zr alloys for alloy design and microstructural prediction. *JOM* 2018;70:2254-9. DOI
31. Bania PJ. Beta titanium alloys and their role in the titanium industry. *JOM* 1994;46:16-9. DOI
32. Hume-Rothery W, Raynor GV. The equilibrium and lattice-spacing relations in the system magnesium-cadmium. *Proc R Soc Lond A* 1940;174:471-86. DOI
33. Zhang YM, Yang S, Evans JRG. Revisiting Hume-Rothery's Rules with artificial neural networks. *Acta Mater* 2008;56:1094-105. DOI
34. Kattner UR. The calphad method and its role in material and process development. *Tecnol Metal Mater Min* 2016;13:3-15. DOI PubMed PMC

35. Tancret F, Toda-Caraballo I, Menou E, Rivera Díaz-del-castillo PEJ. Designing high entropy alloys employing thermodynamics and Gaussian process statistical analysis. *Mater Design* 2017;115:486-97. DOI
36. Wu W, Sun Q. Applying machine learning to accelerate new materials development. *Sci Sin Phys Mech Astron* 2018;48:107001. DOI
37. Wang Y, Wagner N, Rondinelli JM. Symbolic regression in materials science. *MRS Commun* 2019;9:793-805. DOI
38. Cui C, Cao G, Cao Y, et al. Physical metallurgy guided deep learning for yield strength of hot-rolled steel based on the small labeled dataset. *Mater Design* 2022;223:111269. DOI
39. Jiang L, Fu H, Zhang H, Xie J. Physical mechanism interpretation of polycrystalline metals' yield strength via a data-driven method: a novel Hall-Petch relationship. *Acta Mater* 2022;231:117868. DOI
40. Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. *npj Comput Mater* 2018;4:81. DOI
41. Xue D, Balachandran PV, Hogden J, Theiler J, Xue D, Lookman T. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun* 2016;7:BFncomms11241. DOI PubMed PMC
42. Dai D, Xu T, Wei X, et al. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Comput Mater Sci* 2020;175:109618. DOI
43. Rickman JM, Chan HM, Harmer MP, et al. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nat Commun* 2019;10:10533. DOI PubMed PMC
44. Childs CM, Washburn NR. Embedding domain knowledge for machine learning of complex material systems. *MRS Commun* 2019;9:806-20. DOI
45. Murdock RJ, Kauwe SK, Wang AYT, Sparks TD. Is domain knowledge necessary for machine learning materials properties? *Integr Mater Manuf Innov* 2020;9:221-7. DOI
46. Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* 2013;1:011002. DOI
47. Choudhary K, Garrity KF, Reid ACE, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput Mater* 2020;6:440. DOI
48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-30. Available from: https://jmlr.org/papers/v12/pedregosa11a.html. [Last accessed on 25 Sep 2023]
49. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785-94. DOI
50. Abadi M. TensorFlow: learning functions at scale. In: Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming. New York, NY, USA: Association for Computing Machinery; 2016. p. 1. DOI
51. Imambi S, Prakash KB, Kanagachidambaresan GR. PyTorch. In: Prakash KB, Kanagachidambaresan GR, editors. Programming with TensorFlow. EAI/Springer innovations in communication and computing. Cham: Springer; 2021. pp. 87-104. DOI
52. Erickson N, Mueller J, Shirkov A, et al. Autogluon-tabular: robust and accurate automl for structured data. arXiv. [Preprint.] Mar 13, 2020 [accessed 2024 Sep 25]. Available from: https://arxiv.org/abs/2003.06505.
53. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett* 2018;120:145301. DOI PubMed
54. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater* 2019;31:3564-72. DOI
55. Wang H, Zhang L, Han J, Weinan E. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. *Comput Phys Commun* 2018;228:178-84. DOI
56. Wang Z, Han Y, Cai J, Chen A, Li J. Vision for energy material design: a roadmap for integrated data-driven modeling. *J Energy Chem* 2022;71:56-62. DOI
57. Naser MZ, Alavi AH. Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. *Archit Struct Constr* 2023;3:499-517. DOI
58. George EP, Raabe D, Ritchie RO. High-entropy alloys. *Nat Rev Mater* 2019;4:515-34. DOI
59. Ye Y, Wang Q, Lu J, Liu C, Yang Y. High-entropy alloy: challenges and prospects. *Mater Today* 2016;19:349-62. DOI
60. Liu Y, Wang J, Xiao B, Shu J. Accelerated development of hard high-entropy alloys with data-driven high-throughput experiments. *J Mater Inf* 2022;2:3. DOI
61. Rao Z, Tung P, Xie R, et al. Machine learning-enabled high-entropy alloy discovery. *Science* 2022;378:78-85. DOI
62. Wang C, Fu H, Jiang L, Xue D, Xie J. A property-oriented design strategy for high performance copper alloys via machine learning. *npj Comput Mater* 2019;5:227. DOI
63. Zhang H, Fu H, Zhu S, Yong W, Xie J. Machine learning assisted composition effective design for precipitation strengthened copper alloys. *Acta Mater* 2021;215:117118. DOI
64. Ozaki T, Matsumoto H, Watanabe S, Hanada S. Beta Ti alloys with low young's modulus. *Mater Trans* 2004;45:2776-9. DOI
65. Pang C, Jiang B, Shi Y, Wang Q, Dong C. Cluster-plus-glue-atom model and universal composition formulas [cluster](glue atom)$_x$ for BCC solid solution alloys. *J Alloys Compd* 2015;652:63-9. DOI
66. Wang Q, Ji C, Wang Y, Qiang J, Dong C. β-Ti alloys with low young's moduli interpreted by cluster-plus-glue-atom model. *Metall Mater Trans A* 2013;44:1872-9. DOI
67. Jiang B, Wang Q, Wen D, et al. Effects of Nb and Zr on structural stabilities of Ti-Mo-Sn-based alloys with low modulus. *Mater Sci Eng A* 2017;687:1-7. DOI

68. Yang F, Li Z, Wang Q, et al. Cluster-formula-embedded machine learning for design of multicomponent β-Ti alloys with low Young's modulus. *npj Comput Mater* 2020;6:372. DOI
69. Liu X, Tan X. Giant strains in non-textured $(Bi_{1/2}Na_{1/2})TiO_3$-based lead-free ceramics. *Adv Mater* 2016;28:574-8. DOI PubMed
70. Yan B, Cheng L, Li B, et al. Bi-directional prediction of structural characteristics and effective thermal conductivities of composite fuels through learning from finite element simulation results. *Mater Design* 2020;189:108483. DOI
71. Clark SJ, Segall MD, Pickard CJ, et al. First principles methods using CASTEP. *Z Krist Cryst Mater* 2005;220:567-70. DOI
72. Hafner J. *Ab-initio* simulations of materials using VASP: density-functional theory and beyond. *J Comput Chem* 2008;29:2044-78. DOI PubMed
73. Jinnouchi R, Karsai F, Kresse G. On-the-fly machine learning force field generation: application to melting points. *Phys Rev B* 2019;100:014105. DOI
74. Jinnouchi R, Lahnsteiner J, Karsai F, Kresse G, Bokdam M. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with bayesian inference. *Phys Rev Lett* 2019;122:225701. DOI PubMed
75. Noé F, Tkatchenko A, Müller K, Clementi C. Machine learning for molecular simulation. *Annu Rev Phys Chem* 2020;71:361-90. DOI PubMed
76. Zong H, Pilania G, Ding X, Ackland GJ, Lookman T. Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *npj Comput Mater* 2018;4:103. DOI
77. Sun S, Ouyang R, Zhang B, Zhang T. Data-driven discovery of formulas by symbolic regression. *MRS Bull* 2019;44:559-64. DOI
78. Wei Q, Cao B, Deng L, Sun A, Dong Z, Zhang T. Discovering a formula for the high temperature oxidation behavior of FeCrAlCoNi based high entropy alloys by domain knowledge-guided machine learning. *J Mater Sci Technol* 2023;149:237-46. DOI
79. Melhem HG, Nagaraja S. Machine learning and its application to civil engineering systems. *Civil Eng Syst* 1996;13:259-79. DOI
80. Yuan J, Wang Q, Li Z, Dong C, Zhang P, Ding X. Domain-knowledge-oriented data pre-processing and machine learning of corrosion-resistant γ-U alloys with a small database. *Comput Mater Sci* 2021;194:110472. DOI
81. Zou C, Li J, Wang WY, et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Mater* 2021;202:211-21. DOI
82. Tian F, Lin D, Gao X, Zhao Y, Song H. A structural modeling approach to solid solutions based on the similar atomic environment. *J Chem Phys* 2020;153:034101. DOI PubMed
83. Liu G, Sun B. Concrete compressive strength prediction using an explainable boosting machine model. *Case Stud Constr Mater* 2023;18:e01845. DOI
84. Hu Q, Yang R. The endless search for better alloys. *Science* 2022;378:26-7. DOI PubMed
85. Wang Z, Chen A, Tao K, Han Y, Li J. MatGPT: a vane of materials informatics from past, present, to future. *Adv Mater* 2024;36:2306733. DOI