DATA SCIENCE - EXTRACTING INSIGHTS
Open Journal

Review

Small Data Machine Learning in Materials Science

Pengcheng Xu[1]; Xiaobo Ji[2]; Minjie Li[2]; Wencong Lu[1,2,3]*

[1]*Materials Genome Institute, Shanghai University, 200444 Shanghai, China*
[2]*Department of Chemistry, College of Sciences, Shanghai University, 200444 Shanghai, China*
[3]*Zhejiang Laboratory, Hangzhou 311100, China*

****Corresponding author***
Wencong Lu
Materials Genome Institute, Shanghai University, 200444 Shanghai, China; E-mail: minjieli@shu.edu.cn; wclu@shu.edu.cn

Article information
Received: January 21st, 2024; **Revised:** February 24th, 2024; **Accepted:** March 2nd, 2024; **Published:** April 1st, 2024

Cite this article
Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. 2024; 2(1). doi: https://doi.org/10.70705/ppp.dsei.2024.v02.i01.pp58-73

ABSTRACT

This review discussed the dilemma of small data faced by materials machine learning. First, we analyzed the limitations brought by small data. Then, the workflow of materials machine learning has been introduced. Next, the methods of dealing with small data were introduced, including data extraction from publications, materials database construction, high-throughput computations and experiments from the data source level; modeling algorithms for small data and imbalanced learning from the algorithm level; active learning and transfer learning from the machine learning strategy level. Finally, the future directions for small data machine learning in materials science were proposed.

Keywords
Machine learning; Small data; Materials database.

INTRODUCTION

As an interdisciplinary subject covering computer science, mathematics, statistics and engineering, machine learning is dedicated to optimizing the performance of computer programs by using data or previous experience, which is also one of the important directions of artificial intelligence development1,2. In recent years, machine learning has been widely used in many fields such as finance, medical care, industry, and biology3–10. In 2011, the concept of material genome initiative (MGI) was proposed to shorten the material development cycle through computational tools, experimental facilities and digital data. Under the leadership of the MGI, machine learning has also become one of the important means for materials design and discovery11,12. The core of machine learning-assisted materials design and discovery lies in the construction of machine learning models with good performance through algorithms and materials data to achieve the accurate prediction of target properties for undeter- mined samples13. The constructed model could be further used to discover and design materials or explore the patterns and laws hidden behind the materials data. In the past decades, machine learning has become more and more developed and favored by researchers as a powerful tool to assist in the design and discovery of various materials, including alloys, perovskites, polymers, etc14–17. A lot of related studies have proved that compared with the trial-and-error method based on experiment and experience, machine learning can quickly obtain laws and trends from available data to guide the development of materials without understanding the underlying physical mechanism. Data is the cornerstone of a machine learning model, which directly determines the performance of the model from the source. It is widely accepted that we are in an era of big data where the data keep exploding all the time to allow machine learning to play such a big role. However, in the field of materials science, some questions about data are worth thinking deeply. Has the materials data really entered the era of big data? How much data can be considered big data? What is the difference between big data and small data?

Some statisticians consider the 'big' of big data refers to the scale of the data, including the amount of samples or the number of variables18. We believe that the definition standard of big data needs to be determined by combining the sample size and the number of variables. The amount of data needed should vary depending on the size of the space and the complexity of the target system. However, there are few specific quantitative indices about the data size to definite the big data, and there is also obscure to make a clear distinction between big data and small data. The concepts of big data and small data are relative rather than absolute. The small data discussed in this review focuses on the limited sample size. Some scholars believed that the data generally obtained from large-scale observations or instrumental analysis could be regarded as big data, mainly used for simple analysis of prediction; while the data derived from human- conducted experiments or subjectively collection could be regarded as small data, mainly used for complex analysis of

the exploration and understanding of causal relationships18. From this point of view, although the development of materials synthesis and characterization as well as the data storage technology has led to the increase in the amount of materials data, most of the data used for materials machine learning still belong to the category of small data. An important development direction in materials machine learning is the interpretation of the relationship between descriptors and material properties, which can also be viewed as an exploration of causal relationships. However, the applications of the models depend on the accurate prediction ability of the model, so even for small data, there remain some requirements for the prediction ability of the model. The acquisition of materials data requires high experimental or computational costs, leading to the dilemma where researchers must make a choice between simple analysis of big data and complex analysis of small data within a limited cost in the process of data collection. If the goals of the research can be achieved with smaller data, most researchers tend to favor the collection of small samples under the controlled experimental conditions instead of large samples with the unknown origin19. The quality of the data trumps the quantity in the exploration and understanding of causal relationships. In addition, the uncertainty assessment of models constructed with small data is simpler than that of big data, and the conclusions drawn from small data will remind users to use more cautiously. The essence of working with small data is to consume fewer resources to get more information.

Small data tend to cause the problems of imbalanced data and model over fitting or under fitting due to the small data scale and too high or too low feature dimensions, which has always been one of the pain points in materials machine learning. There are two ways to solve the problems caused by small data: One is from the data perspective, to increase the data size in the process of data collection. The other is from the machine learning perspective, to select a modeling algorithm suitable for small datasets or to improve the predictive accuracy of the model through machine learning strategies. As shown in Fig. 1, this review aims to introduce the general process of machine learning- assisted materials design and discovery combined with the cutting-edge research achievements and summarize the methods of dealing with small data in the process. The methods of dealing with small data were introduced from the three levels, including data extraction from publications, materials database construc- tion, high-throughput computations and experiments from the data source level; modeling algorithms for small data and imbalanced learning from the algorithm level; active learning and transfer learning from the machine learning strategy level. In addition, the future directions with challenges of small data in materials machine learning are also summarized.

WORKFLOW OF MATERIALS MACHINE LEARNING

One of the most direct goals of machine learning-assisted materials design and discovery is to apply the algorithms and materials data to construct models for the prediction of the material properties. As shown in Fig. 2, the workflow of materials machine learning includes data collection, feature engineering, model selection and evaluation, and model application20–23.

Materials data are required to be collected after clarifying the research object and relevant properties. The data are generally divided into two parts: The target variable reflecting the property of the materials and the descriptor reflecting the information of the materials themselves. The data of target variable could be collected from published papers, materials databases, lab experi- ments, or first-principles calculations24. Although collecting data from the publications can access the latest research data, it also requires the huge cost to search for a large number of publications along with the data of mixed quality. Besides, even for the same property of the same materials with the same synthesis and characterization methods in different publications, there could still exist some inconsistency in the property values, which may bring the challenges of data uncertainty assessment and the complicated data preprocessing15. A large amount of data can be obtained from the materials databases in a short time. However, due to the cycle delay of the entry and check of the materials data, the data in the latest research could not be available from the materials databases. The quality of data obtained through experiments or calculations tend to be high because of the unification of experimental and computational conditions, but the cost of some materials such as alloys containing precious metal elements is too high to obtain a large amount of data through experiments. The emergence of first- principles calculations has made up for the limitations of experiments. The first-principles method is based on the quantum mechanics, in which the calculation process only requires the involved atomic species and the position coordinates to become one of the preferred methods for the design and exploration of materials25–27. But the calculation accuracy is also affected by the level of material systems and computer hardware. Descriptors can be divided into three scales from microscopic to macroscopic: element descriptors at the atomic scale; structural descriptors at the molecular scale; and process descriptors at the material scale. The element descriptors reflect the composition information of the materials. The acquisition of element descriptors requires the composed chemical elements of the materials and their stoichiometric ratios. Structural descriptors reflect not only compositional information, but also the 2D or 3D structural information of the materials, which can be generated by descriptor generation software or toolkits like Dragon, PaDEL, and RDkit28–31. Process descriptors do not reflect information about the materials themselves, but rather reflect the influence of experimental conditions in synthesis or characterization on the properties. In addition to the above three types of descriptors, generating descriptors based on domain knowledge to construct interpre- table machine learning models is also one of the research hotspots in recent years. Lian et al.32 used machine learning based on domain knowledge to obtain descriptors from empirical formulas containing unknown parameters to predict the fatigue life (S-N curve) of different series of aluminum alloys. Compared with models constructed without domain knowledge, the model predicted ability was greatly improved. For materials data, we have been always insisting that every piece of data is precious and machine learning is able to fulfill its potential value. Descriptors generated from domain knowledge could assist machine learning algorithms to better capture the key information and improve the predicted accuracy of the model.

Feature engineering is an integral part of machine learning. Feature engineering refers to the selection of optimal descriptor subsets from the original descriptors with a series of engineered methods

for modeling, including feature preprocessing, feature selection, dimensionality reduction, and feature combination. Data preprocessing aims to improve the quality of the incomplete, inconsistent, and unusable data. The specific methods include normalization or standardization to perform interval scaling on descriptor data, and convert data with units into data without units to unify data metrics by removing the influence of units to make data processing faster and more agile. For the missing values of the descriptors, the mean, median, before or after values can be used to fill in, or the data corresponding to the missing values can be directly deleted. Materials descriptors, especially those generated by software, tend to be high in the dimension and contain redundant information while describing materials information. The process of removing redundant descriptors is called feature selection. According to the relationship between the feature selection algorithms and modeling algorithms, the commonly used feature selection methods can be divided into filtered, wrapped and embedded33,34. In addition to the feature selection, the descrip- tors of the original high-dimensional space can also be reorganized to reduce the dimensionality by projecting the descriptors of the original high-dimensional space into the low- dimensional space, which is called dimensionality reduction35,36. The difference between dimensionality reduction and feature selection is that feature selection aims to remove and delete the redundant descriptors, while dimensionality reduction is to form descriptors through the reorganization of descriptors and does not retain any of the original descriptors. Common dimensionality reduction methods include principal component analysis (PCA) and linear discriminant analysis (LDA)37–39. Feature combination could deal with the problem of under fitting caused by too low descriptor dimensions. The core of feature combination is to generate a lot of combined descriptors by combining the original descriptors with the simple mathematical operation for further feature selection and modeling. The Sure Independence Screen- ing Sparsifying Operator (SISSO) is a compressed sensing-based data analysis method that can perform feature engineering transformations based on given descriptors to generate a large number of features, from which the optimal low-dimensional feature subset could be found40,41.

There are various modeling algorithms to choose for either regression or classification tasks. For the same data, models constructed with different machine learning algorithms have different performance, which requires the evaluation of the modeling algorithms to select the optimal model without any under fitting and over fitting. The most used evaluation methods are K-fold cross-validation (K-fold CV), leave-one-out cross- validation (LOOCV), and leave-out method42–44. K-fold CV randomly divides the original data into K parts by non- repetitive sampling and selects 1 part as the test set each time, while the remaining K-1 parts are used as the training set for modeling. After repeating K times, total K models are obtained after training on each training set to test the performance with the corresponding test set. The average of the K-group test set results is used as a performance indicator to evaluate the model performance under the K-fold CV. LOOCV is a special case of K-fold CV, where K is equal to the number of samples N. Therefore, for N samples, there are N-1 samples selected each time to train the model, leaving one sample as the test set to evaluate the model. The leave-out method refers to dividing the original dataset D into two mutually exclusive subsets S and V. The training set S is used to train the model, while the test set V is sets the unknown data used to evaluate the generalization ability of the model. It should be noted that the K-fold CV and the leave- out method have certain requirements on the data size. Especially when the number of samples is less than 30, LOOCV is generally considered to be the most recommended evaluation method. In case the division of the dataset may have an impact on the performance of the model, the repeatability measure named y-scrambling can be used to further verify the stability of the model45,46. By randomly dividing the dataset into training set and test set for multiple times to evaluate the stability of the model, the problem of random fluctuations caused by dataset division can be avoided. After the evaluation method is determined, specific indicators are needed to quantify the performance of the model. For regression tasks, commonly used evaluation indicators include mean absolute error (MAE), mean relative error (MRE), root mean square error (RMSE), correlation coefficient (R), and the determination coefficient (R2) between the predicted value and the true value. For classification tasks, commonly used evaluation indicators include classification accuracy, true positive rate (TPR), false positive rate (FPR), recall rate, precision rate, etc. In model selection and evaluation, it is necessary to consider the influence of algorithm parameters on the model. The process of parameters optimization aims to adjust the model parameters to further improve the prediction ability of the model.

The most basic function of a model is to predict the properties of the unknown materials. According to this function, the model can be applied to virtual screening, online server and theoretical discovery. Virtual screening refers to artificially generating a large number of virtual samples for the constructed models to predict properties and quickly screen out the materials that meet the requirements for further experimental or computational valida- tion47,48. Virtual screening avoids the experience-based experi- ments to a certain extent and realizes the data-driven way to design and discover materials. However, the generated virtual samples often cannot cover the entire search space and huge computing resources are still consumed in the prediction of too many samples. The online server allow the constructed models to be imported into the back-end server and then the corresponding user interaction page is developed on the front-end49. Before researchers conduct experiments of the designed materials, the properties can be quickly obtained through model prediction once the user inputs the necessary information for unknown samples. The advantage of the online server lies in the sharing of models, where the models can be used anytime and anywhere with only electronic equipment and network. Both virtual screen- ing and online server are the most intuitive applications of the model without any exploration of the laws and patterns contained in the materials data. While the theoretical discovery could explore the relationship between the important material descriptors and properties with the assistance of statistics and domain knowledge to better understand the nature of the materials properties and guide the design of materials. However, we should be cautious when using the rules mined from small datasets because the rules are only more suitable for small data and the generalization ability remains to be verified.

INCREASE THE DATA SIZE BEFORE/IN THE DATA COLLECTION

In this part, some methods for small data in materials machine learning before/in data collection will be introduced with the combination of cutting-edge and typical cases. As has been illustrated above, the materials data tend to be collected from publications, materials database, first-principles computations and experiments. Therefore, data extraction from publications, materi- als database constructions as well as high-throughput computa- tions and experiments could help obtain the data size from the data source.

Data extraction from publications

The publications often contain data of the most cutting-edge studies. Most of the data collected from publications in the materials machine learning work rely more on the human resources to search and read publications for data collection. The most famous inorganic crystal database, the database of materials platform for data science (MPDS) created by the team of Villars, obtains the data by manual review of publications before entering the data into database15. Nevertheless, manually extracting data from publications is rather expensive and labor- intensive. In addition, in the process of manual data collection, the bias of data caused by subjective factors would occur, leading to the situation where the data not conducive to modeling tend to be ignored or directly removed. The model construction in the bias of the data is extremely detrimental to the model applications. With the development of natural language proces- sing (NLP) and text mining (TM) technology, the ideal of automatic data extraction from publications is expected to be realized50,51. The commonly used software and platform in TM could be available in Supplementary Table 1.

The main steps of automatically extracting data from publica- tions through NLP and TM technology include: (1) document retrieval and conversion into plain text; (2) text preprocessing, including sentence labeling and segmentation, text normalization, part-of-speech labeling and dependency parsing; (3) information retrieval; (4) data management52,53. The retrieval of documents mainly refers to the search of published papers in different journals. However, many journal papers are not open access and require plenty of money to subscribe. Besides, the format and layout of papers in different journals vary a lot, leading to the barriers in automatically data extraction. In addition to journal articles, documents such as conference papers, patents, technical reports, etc. also contain the required information. Documents in journals are mostly in the form of HTML or PDF files. The HTML can be parsed and marked up with programming tools, while the PDF files are complex in the form where the arrangement of text is interspersed with tables, figures and equations, which affects the accuracy of conversion to original text and increases the difficulty of extracting plain text from PDF files. In the process of converting PDF documents, errors often occur due to the superscripts and subscripts in chemical formulas or equations. Such errors require advanced optical character recognition (OCR) to avoid54. Creating an OCR for scientific texts is an area of active research in computer science and TM. The labeling and segmentation of sentences is the key step in information extraction to better understand the logical components in sentences. The labeling of sentences requires the explicit labeling criteria, usually marked with special symbols; while the segmentation of sentences aims to determine the boundaries in the text55. However, the complexity of materials terminology and non-standard naming conventions in academic papers often lead to labeling errors and over-segmentation of sentences, which would propagate along the TM process to affect the accuracies of results. Text normalization can be understood as stem extraction56. The same word usually has different existence in different tenses and voices. Extracting the stem of the word would help to reduce the complexity of language. Part-of-speech (POS) labeling refers to identifying and marking the grammatical properties of words, such as verbs, nouns, adjectives, etc., which is used to provide the language- and grammar-based lexical features to TM models57. But in scientific texts, the ambiguity caused by the context of words brings challenges to POS labeling and requires adjustments to the underlying NLP model. Dependency parsing could map linear sequences of sentence tokens to hierarchies by parsing the internal grammatical dependencies between words, which is highly sensitive to the accuracy of punctuation marks and word forms58. In scientific papers, to describe the objectivity of facts, the authors tend to use a lot of passive voice and past tense, resulting in that the general dependency parsing models cannot accurately capture the features of the sentences. Information retrieval (IR) refers to the use of NLP techniques to extract various types of data from the preprocessed text, of which the most common IR method is named entity recognition (NER), which classifies text tokens into specific categories59,60. In scientific texts, named entities can be technical terms as well as physicochemical parameters and properties. Chemical NER is a widely used IR method that usually involves the identification of chemical and materials terms in the text with early applications focusing on the extraction of drug and biochemical information61. Data in academic papers exist not only in text, but also in figures and tables that are embedded in the text. Extracting data from journal figures and tables requires both TM and image recognition techniques. The challenges to data retrieval caused by the format of figures and tables in academic papers include: (1) Figures and tables exist not only in text, but also in external links such as the supporting information. (2) The forms of figures and tables are very complex. For example, the figure could be mixed with a table; the figure could contain multiple sub-figures; and the table row and column could be merged. Although image recognition technology has been widely used in materials science, it is more used to explore the morphology and structure of the materials in figures through deep learning, rather than to separate the figures embedded in scientific texts.

At present, manual data extraction from publications is still the mainstream. The ambiguity of materials naming standards, the complexity of the chemical formulas, the diversification of languages, and the professional terminology have all caused great challenges to apply NLP and TM technology to automatically extract data from publications. Although automated data extrac- tion from publications is still in its infancy, TM and NLP may play a key role in enabling more data-driven materials research. Swain et al.62 has developed the toolkits called ChemDataExtractor for automatic extraction of chemical information from scientific publications. ChemDataExtractor provides a layout analysis tool for complex PDF files built on the PDFMiner framework to group text into headings, paragraphs and captions using the position of images and text characters. Besides, ChemDataExtractor could group text into headings, paragraphs and captions using image and text character positions. For text labeling and segmentation, ChemDataExtractor provides a sentence splitter using the Punkt algorithm based on Kiss and Strunk, which detects

sentence boundaries through unsupervised learning of common abbrevia- tions and sentence beginnings. The Punkt algorithm has been proved to be widely applicable to multiple languages and text domains, performing the best when trained on text from the target domain. For words derived from unannotated publications, Brown clustering is used to implement hierarchical clustering based on the context of word occurrence to improve the performance of lexical labeling and named entity recognition in various domains. The POS tagger of ChemDataExtractor is trained with a linear chain conditional random field (CRF) model using the orthant-wise limited-memory quasi-Newton (OWL-QN) method implemented in the CRFsuite framework. A CRF model-based recognizer combined with a dictionary-based recognizer and a regular expression-based recognizer are used for chemical named entity recognition. For chemical identifier disambiguation, the Hearst and Schwartz algorithms are used to detect the definition of chemical abbreviations and labels, which could generate a list of mappings between abbreviations and their corresponding full non-abbreviated names, merging the data defined for different identifiers into a single record for each chemical entity. In addition to extracting data from text, ChemDataExtractor can also parse tables to extract data. For tables where each row corresponds to a single chemical entity and each column describes the value of that entity's attributes, ChemDataExtractor can extract information using a dedicated version of the natural language processing pipeline by treating each individual table cell as a short, highly formulaic sentence. Currently, the team has released ChemDa- taExtractor version 2.0, which retains all the features of ChemDataExtractor while providing a complete approach to ontology auto-population in the scientific domain63. ChemDataEx- tractor 2.0 supports extraction from publications from 155 papers as an evaluation set, using extracted data from each compound with 18 sets of nested crystallographic features, which generated an overall accuracy of 92.2% across 26 different journals, achieving the construction of a framework for seamless integration from publications to data-driven methods. Yukari et al.64 developed a web-based system called Starrydata2 to automatically extract numerical data from figures of scientific papers and the chemical composition of the corresponding samples. The visualization capabilities of Starrydata2 allow for the display of data files in a variety of formats, including line plots, heat maps, and multiple scatter plots. Starrydata2 has successfully collected experimental data from mapped figures of more than 11,500 samples of thermoelectric materials. The electronic structure differences of the parent compounds PbTe, PbSe, PbS, and SnTe were revealed by combining a partial experimental dataset of 434 rock salt-based thermoelectric materials with first-principles calculations. The evaluation of the electronic relaxation time τel by combining the computational and experimental data revealed that achieving a long τel is considered essential to improve the thermoelectric quality factor.

Materials database construction

Materials data have the characteristics of high reliability requirements, strong correlation, many influencing factors, complex acquisition process and wide distribution of data, which is one of the reasons for the dilemma of small data in materials science. The materials database could collect the fragmented materials data conveniently for users to store, update and retrieve large amounts of data more quickly, safely and accurately. In the design and discovery of materials with data-driven methods, the acquisition of the materials properties, the mechanisms under special conditions, materials performance improvement, materials selection and safety evaluation are all inseparable from the support of materials database platforms. Obtaining a large amount of materials data through databases for further analysis and knowledge mining is one of the important directions of materials machine learning. The most common way to use the data in the database is to be taken as the training set to train the machine learning model combined with the algorithms. In addition to the training set, the data in the database can also be used as a test set to evaluate the performance of the constructed model, or used as a candidate set in combination with the model to filter out the materials with the properties meeting the requirements. Many databases in recent years tend to have high-throughput computing frameworks, machine learn- ing toolkits, and statistical analysis tools, which indirectly provide support for machine learning research.

The commonly used materials databases are shown in Supplementary Table 2. According to the data types in the materials databases, materials data can be divided into computa- tional and experimental data. Computational data refer to theoretical data on materials, usually derived from high- performance and high-throughput computations based on first principles. It should be noted that the computational data need to be combined with experimental data and empirical data to process and analyze large-scale materials data to be fully mined and utilized. Most of the experimental data mainly exist in the publications or the private database where the researchers could enter the data after the experimental synthesis and characteriza- tion. Both computational and experimental data could be automatically extracted from publications through NLP and TM techniques. The 35,675 solution-based databases of inorganic materials synthesis procedures were extracted from over 4 million publications using NLP, TM and machine learning by Ceder et al.65 Each of these procedures contains the basic synthesis information, including the parent ion, target materials, quantities, synthesis action and corresponding properties. The experts verified the completeness and accuracy of the data by randomly extracting data combined with domain knowledge. In addition, the diversity of the extracted data was further analyzed in relation to the spatial extent of the materials covered. The results of the analysis show that the common targets and their corresponding precursors in the dataset cover materials that have attracted extensive attention over the last two decades. The database contains a large-scale solution-based dataset of inorganic material synthesis procedures, providing a basis for testing and validating the existing empirical synthesis rules, improving prediction accuracy, and even mining rules to guide synthesis. For both the computational data and the experimental data, the most intractable difficulty in the construc- tion of a material database is the evaluation and verification of data quality. Although there are many materials databases, each one has its own standards for the evaluation and verification of data quality, which are not uniform. Even though many scholars are working on developing material data standards, there is few specific standard for the evaluation and verification of materials data quality66. Therefore, it is necessary to learn the research experience of data quality evaluation in other fields, combining the characteristics of materials data to carry out research on the data quality evaluation methods, corresponding managements and applications. It can be

found from Supplementary Table 2 that many materials databases are established according to the types of materials, but the classification of materials can be divided into many types according to the different standards. The obscure classification standards of materials also bring obstacles to the construction of material databases. Materials can be defined by their chemical composition and structure, while most databases use only chemical composition or chemical formula to identify materials, which could cause the situation where the materials of different structures are often indistinguishable. Xu et al. proposed the MatML, a specification designed for material information exchange, which uses chemical composition and processing conditions to describe materials, based on research experience on materials such as single crystals, ceramics, alloys, polymers, etc. and the basics of materials science67. Materials could be divided into four levels according to MatML in Fig. 3: chemical system, compound, substance and material. Chemical systems are the basis of all materials to represent one or more elements that make up a material. Compounds are the second level to identify materials at the molecular level. For most inorganic materials, a compound can be defined by the chemical formula. However, for organic or polymer materials, the molecular structure must be specified. The third level is substance, which determines the state of the compound such as gas, liquid or solid. For solid state, the crystal state and crystal structure should also be given. A substance should correspond to a phase in a phase diagram. The fourth level is the materials. To define a material, many types of information are required, such as the form, dimensions, microstructure, process conditions, etc. In addition, the polymer database is extremely limited in the materials databases, which may be due to the structural properties of polymer materials. Experimentally synthesized polymers are often rarely single entities. The same polymer material with different polymerization degrees leads to different molecular weight distributions, which has created great difficulty to the database construction. Also, the complex monomeric structures and sequences of polymers lead to the lack of standard naming rules. The challenges currently faced in the construction of polymer databases include appropriate descriptors, elements associated with properties, details of characterization, and sources of data68.

The materials database has developed for decades to store the materials data, but the construction of the materials database still faces great challenges. Firstly, there are many scientific research institutions establishing the materials databases in the world, which leads to the fragmentation of materials databases and low utilization rate of the materials data due to the uneven data quality of the database. In addition, the materials database is still in the stage of simultaneous construction and use. The construc- tion and maintenance of the database have been the long-term work, requiring a lot of capital and human resources as well as the professional supervision for the collection, update and main- tenance of the data. Secondly, the lack of uniform and complete materials classification standards and data quality evaluation methods would lead to uneven data quality. Lastly, the sharing of the materials data is limited due to intellectual property issues. In the establishment of the materials databases, the management of the intellectual property rights of the data should be strength- ened; the quality and sharing of the data in the database should also be improved69. Even though the construction of the material database faces many challenges, the rapid acquisition of data from the materials database has alleviate the problem of small data to become an important way of data collection for materials machine learning.

High-throughput computations and experiments

Materials data are rather precious because of the high experi- mental and computational costs. But the presence of high- throughput technology makes it possible to obtain a large amount of high-quality data by experimental or computational methods in a short period of time. The concept of high-throughput stems from gene sequencing. The first-generation sequencing can only measure one sequence of one sample at a time to generate relatively small data, while high-throughput sequencing can measure a large number of samples at a time, resulting in data in the dozens of gigabytes even hundreds of gigabytes70,71. One of the characteristics of high-throughput lies in the ability of processing a lot of samples in a short time to obtain more data. With the development of first-principles, high-performance computers and materials preparation and characterization tech- nologies, obtaining a large amount of high-quality materials data through high-throughput computations and experiments combined with machine learning to develop materials is also a solution to small data in materials machine learning.

First principles are the cornerstone of high-throughput computations, which enables accurate computation of various electronic structures and total energy-related properties under atomic structure, including the properties of thermodynamics, kinetics, electromagnetism, and mechanics72,73. The development of high- performance computer and computational simulation technology makes the first-principles-based high-throughput computations to screen materials a potential direction in the materials design and discovery. Prior to experimental design, high-throughput computations can be used to screen out the stable structures that meet the requirements. The commonly used high-throughput computation toolkits are shown in Supplementary table 3. The essence of materials design based on high-throughput computa- tions is to apply the concepts of 'blocks construction' and 'high- throughput screening' in combinatorial chemistry to the computer simulation of materials. After determining the basic building blocks of composition through materials calculations, a large number of compounds are constructed to obtain the correspond- ing properties through high-throughput computations, where machine learning could integrate data, program, and materials calculation software to map the quantitative relationship models of material composition, structure, and properties to guide the design of materials. As shown in Fig. 4, the workflow of the materials screening by high-throughput computations is generally divided into five steps: the construction of the samples for high- throughput computations and screening; screening based on thermodynamic stability; preliminary screening based on basic descriptors with limited precision; specific screening based on high-precision descriptors; screening based on other conditions74. High-throughput computations use the density functional theory (DFT) methods to quantitatively or qualitatively calculate the relevant properties of a large number of initial input material structures for screening, while machine learning combines the data to construct models to explore the patterns and laws behind the data. Both high-throughput computations and machine learning could essentially extract valuable information from data.

However, high-throughput computations are more inclined to complete the specified work according to the set rules such as calculating the properties of materials according to first-principles methods, which do not have the generalization ability. While machine learning tends to perform the good generalization ability because of the decision-making nature of the modeling algo- rithms. Combining high-throughput computations with machine learning to fully take the advantages of the parameters standardization and large-scale of high-throughput to solve the problem of small data in machine learning is expected to further improve the efficiency of screening and development of materials. High-throughput experiment, also known as high-throughput preparation and characterization technology, is an important part of MGI75. The core idea of high-throughput experiment is to change the original sequential iteration method into parallel or efficient serial experiments. Commonly used high-throughput preparation and characterization techniques are shown in Supplementary table 4. The high-throughput preparation of materials is also called the combined preparation of materials, which refers to the preparation of a large number of materials with different components in a short time by a certain experimental method. After the materials preparation, high- throughput characterization techniques are required to obtain sample information in a relatively short time for further experiments or detailed characterization. The materials design processes of traditional way and the high-throughput experiments are shown in Fig. 575. The traditional materials design includes the loop of experimental design, material synthesis/characterization, and materials property analysis. Compared with traditional methods, the materials design based on high-throughput experi- ments takes the database as the center of the loop, integrating the data collection, storage, management, and mining to make full use of data to promote the development and applications of materials. High-throughput experiments can rapidly accumulate a large amount of experimental data to facilitate the screening or optimize the applications of materials.

High-throughput computations and experiments have become significant methods to provide sufficient materials data for machine learning research. Hu et al.76 obtained 640 2D halide perovskites A2BX4 (A = Li, Na, K, Rb, Cs; B = Ge, Sn, Pb; X = F, Cl, Br, I) and corresponding adsorption energies with Li+, Zn2+, K+, Na+, Al3+, Ca2+, Mg2+, and F− by using high-throughput computations. After filtering out 13 descriptors with the Pearson correlation coefficient, k-nearest neighbors (KNN), Kriging, Random Forest, Rpart, SVM, and XGBoost were adopted for modeling. The results revealed that XGBoost performed the highest prediction accuracy with the R2 and RMSE of the training set being 0.998 and 0.128 eV, respectively. After modeling, various methods were used to rank the importance of descriptors, and different ranking methods consistently showed the great importance of ionic adsorbent density on the adsorption energy of hybrid systems. After high- throughput screening, 5 candidates were screened from a virtual design space consisting of 11,976 ion/perovskite for DFT verification, which proved to be applicable to ion batteries. Hayashi et al.77 developed an open-source Python library named RadonPy for fully automated polymer property calculations using all-atom classical molecular dynamics (MD) simulations, and successfully performed high-throughput computations on more than 1,000 amorphous polymers with a wide range of thermo physical properties. Machine learning techniques were success- fully applied to calibrate the bias and variance of MD calculations. 8 amorphous polymers with high thermal conductivity and the underlying mechanisms were identified after high-throughput screening by RadonPy. The construction of a database using RadonPy will rapidly yield a large amount of high-quality data on polymer properties to facilitate the development of polymer informatics. Zhao et al.78 explored the optimal stability conditions for organolead iodide perovskite cells using a high-throughput experiment-based robotic system and machine learning. The robotic system synthesized more than 1,400 perovskite battery samples under different material compositions, experimental conditions, test conditions, and measured the battery perfor- mance decay time as the battery stability standard. Taking the material composition, experimental conditions and test conditions as the descriptors, and the battery performance decay time as the target property, a gradient boosting tree model was constructed with the RMSE of the test set being 169 h. According to the optimal experimental conditions obtained from the feature analysis and the optimal composition of the optimal organic lead-iodine perovskite MA0.1Cs0.05FA0.85PbI3, a perovskite battery with the highest performance degradation time of more than 4000 h was successfully synthesized, far exceeding the vast majority of reported battery devices. This work has obtained the effect of optimal experimental conditions on battery performance degradation through high-throughput experiments and machine learning analysis, which effectively promotes the progress of perovskite battery stability research.

Both high-throughput computations and experiments in the above researches are providing the sufficient sample size for modeling. However, since high-throughput computations are developed based on first-principles calculations, the computa- tional characterization also brings more potential application possibilities in combination with machine learning. Machine learning can also be used to improve the precision and accuracy of DFT calculations. James et al.79 trained a neural network called DeepMind21 (DM21) on molecular data and fictitious systems with fractional charges and spins to overcome systematic errors due to violations of the mathematical properties of exact generalized functions. DM21 provides a solution to the accuracy and precision problems associated with DFT calculations, demonstrating the success of combining DFT with modern machine learning methods. For different DFT computational data, it is often necessary to construct different machine learning models to ensure model accuracy. Developing machine learning models with general applicability to different DFT data is also one of the current research directions for combining machine learning with DFT computations. Takamoto et al.80 trained a generalized neural network potentials (NNPs) model called prefiring potentials (PFP) using 20 datasets of DFT calculations. PFP is capable of handling any combination of 45 elements and has general applicability in different application fields, including lithium diffusion in LiFeSO4F, molecular adsorption in metal-organic frameworks, anorder–disorder transition of Cu-Au alloys, and material discovery for a Fischer–Tropsch catalyst. PFP can greatly alleviate another limitation of atomic simulations caused by time and space scales. The combination of DFT and PFP or experiments using PFP-based screening will also accelerate the field of materials discovery.

ALGORITHMS FOR SMALL DATA IN MODELING

In the modeling process, some algorithms have good compat- ibility with small datasets and unbalanced data to obtain the ideal results. This part will introduce small data modeling algorithms and algorithms for dealing with imbalanced data.

Modeling algorithms for small data

The performance of machine learning models is not only dependent on the quantity and quality of the data, but also highly dependent on the modeling algorithm. Some algorithms are well appropriate for modeling with small data. Combined with the case study of materials machine learning, algorithms suitable for modeling with small data include support vector machine, Gaussian process regression, random forest, gradient boosting decision tree, XGBoost and symbolic regression.

Support vector machine (SVM) is a kernel-based algorithm. The kernel functions would efficiently complete the space transforma- tion to convert the original nonlinear problem into a linear problem in a high-dimensional space and turn a linear inseparable problem in a low-dimensional space into linearly separable81. The basic principle of the SVM is to map the input vectors to a high- dimensional space, finding an optimal hyperplane as the criterion for sample classification to achieve the best compromise between model complexity and learning ability to obtain the best robustness82. According to the machine learning tasks of classification and regression, SVM can also be called support vector classification (SVC) and support vector regression (SVR). The goal of SVC is to obtain the classification line of the largest edge hyperplane, where samples of different classes could be the farthest from each other. SVR uses the insensitive channel ε to deal with the trade-off between empirical risk and structural risk. The error is ignored when the predicted value ŷi satisfies $|y_i-\hat{y}_i| \leq \varepsilon$, otherwise the error is $|y_i-\hat{y}_i|-\varepsilon$. In the empirical risk calculation, only the deviation is considered when it is greater than ε. The concept of "margin" in SVM has provided a structured description of data distribution, thereby reducing the requirements for data size and data distribution.

Gaussian process regression (GPR) is a non-parametric method with Gaussian Process (GP) priors to perform regression analysis on data83. The model assumptions of GPR include regression residuals and Gaussian process priors. Without restricting the form of the kernel functions, GPR is theoretically a universal approximator for any objective function in the compact space. In addition, GPR could provide the posterior distribution of the predicted result with an analytical form when the regression residuals are normally distributed, which has proved that GPR is a probabilistic model with generalization ability and interpretability. As a non-parametric Gaussian process model, the complexity of GPR depends on the training data. Based on the characteristics of Gaussian process and the kernel functions, GPR is usually used for regression modeling of low-dimensional and small data.

Random forest belongs to the Bagging type ensemble algorithm. By combining multiple weak classifiers, the final result is obtained by voting or average to improve the prediction accuracy and generalization performance of the overall model. A random forest consists of multiple decision trees and each tree in the forest jointly determines the final output of the model84. First, bootstrap sampling is applied to randomly select k samples from the original training set with replacement to form training samples. Then, the models of k decision trees are constructed for each of the k samples to randomly combine to form the random forest. Finally, each record is voted to determine the final classification according to the k classification results. For classification tasks, each decision tree in the random forest will give the final category, and finally the output category of each decision tree in the forest is comprehensively considered by voting. For regression tasks, random forest takes the average output of each decision tree as the final output.

Gradient boosting decision tree (GBDT) is an iterative decision tree algorithm consisting of multiple decision trees that generate multiple weak learners in series85. By fitting the negative gradient of the loss function of the previous accumulated model of each weak learner, the accumulated model loss after adding the weak learner is reduced in the direction of the negative gradient. Each tree can make predictions on part of the data to get the final result by adding the conclusions of all the trees. Gradient boosting can be used for both classification and regression tasks.

XGBoost is an efficient system of Gradient Boosting, which realizes to form the tree with the difference between the result of the basic learner and the actual value to reduce the difference between the model value and the actual value to avoid over fitting86. When using classification and regression trees (CART) as the base classifier, XGBoost explicitly adds a regular term to control the complexity of the model, which is beneficial to prevent over fitting and improve the generalization ability of the model. GBDT only takes the first-order derivative information of the loss function during model training, while XGBoost performs second- order Taylor expansion on the loss function to use both the first- order and second-order derivatives at the same time. The traditional GBDT uses CART as the base classifier, while XGBoost supports multiple types of base classifiers, such as linear classifiers. Traditional GBDT uses all the data in each iteration, while XGBoost adopts a strategy similar to random forest, which supports sampling of data. The traditional GBDT is not designed to deal with missing values, while XGBoost can automatically learn the processing strategy of the missing values.

Symbolic regression is a genetic programming-based machine learning technique designed to identify an underlying mathe- matical expression87,88. It first builds a stochastic formula to represent the relationship between known independent and dependent variables to predict data. Each successive generation procedure evolves from the previous one, selecting the most suitable individuals from the population for genetic operations such as crossover, mutation and reproduction. The mathematical expression generated by symbolic regression is a combination of operator functions, variables and constants, which is essentially a combinatorial optimization process based on symbolic sets and intelligent algorithms. Currently, symbolic regression has been widely used in the field of materials machine learning to explore the relationship between important descriptors and material properties as well as to construct interpretable machine learning models.

Some of the works have proved that some algorithms have ideal performance in small data modeling. Weng et al.89 used symbolic regression to design a simple descriptor for describing and predicting the oxygen evolution reaction (OER) activity of oxide perovskite catalysts to rapidly identify oxide perovskite catalysts with improved

OER activity. 18 known perovskite catalysts were first synthesized experimentally, 4 samples of each. Each sample was subjected to 3 OER tests under the same conditions and the reversible hydrogen electrode voltage (VRHE) was measured at 5 different current densities, resulting in 1080 data points. The electronic parameters such as the number of d electrons for TM ions, electronegativity values χA and χB, valence states QA, ionic radii RA, the tolerance factor t and the octahedral factor μ were combined with symbolic regression and hyper parametric grid search to generate about 8,640 mathematical formulas. After evaluating the accuracy and complexity of the generated formulas, the 9 mathematical formulas at the Pareto front meet the criteria of high accuracy and low complexity, with the descriptor of μ/t being the best compromise between complexity and accuracy. The descriptor of μ/t is able to reveal the pattern between the OER activity of oxide perovskite catalysts and the structural factors. Smaller μ and larger t would lead to higher OER activity, so the use of large cations at the A-site and small cations at the B site of the perovskite structure enable further development of a large number of previously unexplored OER catalysts. After screening 3545 oxide perovskites in combina- tion with virtual screening, 13 samples with minimum μ/t values were selected for experimental validation. The experimental results show that 5 pure oxide perovskites possess OER activity, with Cs0.4La0.6Mn0.25Co0.75O3, Cs0.3La0.7NiO3, SrNi0.75Co0.25O3 and Sr0.25Ba0.75NiO3 exhibiting OER activity exceeding that of oxide perovskite catalysts reported in the publications. Shi et al.90 collected 50 ABO3-type perovskites and corresponding experi- mental specific surface area (SSA) values from the publications as target property, 40 of which were used as training set and 10 as test set. The descriptors of atomic parameters and sol-gel process parameters are combined with genetic algorithm (GA) and SVR to select the optimal feature subset and construct the model for SSA prediction. The RMSE values of the training set and test set of the model are 3.745 and 1.794 m2 g−1, respectively, indicating the high prediction accuracy of the model. In addition, sensitivity analysis was used to analyze the quantitative impact of 5 important descriptors on SSA and 5 candidates with higher SSA were screened out by virtual screening. The author also developed a web server to realize real-time sharing of the model, laying a foundation for machine learning-assisted design of ABO3-type perovskites with high SSA. Lu et al.91 collected experimental interlayer spacing data for 85 layered double metal hydroxides from publications, 68 of which were used as training set and 17 as test set; and atomic parameters were collected from Lang's handbook of chemistry as descriptors. The algorithms of GA combined with XGBoost, SVR and artificial neural network (ANN) were adopted to select features and construct the model. It is found that the XGBoost model with 6 descriptors performs the best. After randomly splitting the dataset 4 times, the average R of LOOCV and test set could reach 0.91 and 0.87, respectively. After parameters optimization, the LOOCV and test set R values of LOOCV are as high as 0.94 and 0.89. After virtual screening with the constructed model, Co0.67Fe0.33[Fe(CN)6]0.11•(OH)2 with the interlayer spacing up to 12.4 Å was screened out to applied to super capacitors.

In addition to modeling algorithms for small data, our team have integrated various algorithms through ensemble learning to improve the predicted accuracy of the model constructed with small data. Chen et al.92 proposed a step-by-step design strategy based on small data to aid in the design of low melting point alloys. Ridge regression, XGBoost and SVR were applied to screen out three sets of optimal feature subsets and respectively constructed the melting point prediction models of low melting point alloys. After evaluating model performance with 10-fold CV, it was found that the performance of the three models was similar, and the R of the models were all higher than 0.94. In order to obtain a model with more stable prediction ability and higher accuracy, the R-X-S (Ridge regression-XGBoost-SVR) ensemble model was obtained through arithmetically integrating the three models by taken the average value. The R of the R-X-S model in the test set reached 0.990, which was higher than the highest single model. As shown in Fig. 6a, in order to further verify the generalization ability of the model, an external validation set was used to verify the four models, and the R of the four models were all higher than 0.97, which has proved that the R-X-S model has strong generalization compared with 0.968 of the original value in the publication. Besides, Lu et al.93 carried out a study on predicting the bandgaps for hybrid organic-inorganic perovskites (HOIPs) by using ensemble learning. The authors collected 1201 samples from the publications from 2009–2021, and generated 129 atomic descriptors, including atomic radius, atomic chemical potential, tolerance factor, tau factor, octahedral factor, etc. Then the various modeling algorithms were adopted to construct the models. And the top 4 models, involving CatBoost, XGBoost, LightGBM and gradient boosting machine (GBM) were selected as the sub-models for the ensemble learner namely, the weighted voting regressor (WVR). The WVR model complimented the weakness of each sub-model, and achieved a comprehensively superior performance than the sub-models. The R2 and RMSE in LOOCV of WVR reached 0.95 and 0.079 eV respectively, while the R2 and RMSE in test set of WVR achieved 0.91 and 0.106 eV respectively. Based on the ions collected from the formulas of the dataset, the authors constructed a gigantic material space compromising over 8.2 × 1018 combinations for exploring HOIP structures with suitable bandgaps. The proactive searching progress (PSP) method was developed to efficiently search the material compositions with expected bandgap values from the universal chemical space. As the result of PSP method shown in Fig. 6b, the 20,242, 733,848, 764,883, and 746,190 non-Pb samples were designed for the HOIPs with the bandgaps of 1.20 eV, 1.34 eV, 1.70 eV and 1.75 eV, respectively. To validate the searching result of PSP method as well as the predicting ability of the WVR model, the HOIP components of MASnxGe1-xI3 (x = 0.85, 0.74, 0.66) were synthesized and characterized as the experimental validation, where the average error between experiments and predictions was only 0.07 eV. The data in Lu's work have reached up to 1201 and may be far from small data. But the constructed ensemble model with the superior performance to the sub-models indeed indicates that integrating various algorithms through ensemble learning could improve the predicted accuracy of the model.

Imbalanced learning algorithms

Imbalanced learning algorithms aim to deal with the imbalanced data caused by the small data in the classification. Imbalanced learning is aimed at classification tasks, which is mainly manifested in that data size in different categories is unbalanced due to the limited samples in the minority class94. The minority samples of unbalanced data can be divided into absolutely and relatively few in terms of data

size95. Absolutely few data refer to the data size of the minority class itself is rather scarce to lead to the limited information contained in data, which would be difficult for the classifier to capture the information of the minority class samples. Relatively few data mean that the minority class samples only occupy a small proportion compared with the majority class samples to blur the boundary of the minority class sample and reduce the recognition ability of the minority class samples. Traditional classification methods usually process data when the data size of each category is almost equal, but the data categories in materials science are often unbalanced.

Imbalanced learning aims to deal with imbalanced data from two levels of data preprocessing and algorithm. The introduction of the commonly used imbalanced learning algorithms are available in Supplementary table 5. The most basic data preprocessing method is sampling, including undersampling, oversampling, and mixed sampling95. Undersampling balances the minority class by reducing the number of majority class samples, while oversampling by increasing the number of samples in the minority class to balance the data. Mixed sampling combines the oversampling and the undersampling to balance the data size of different categories. Algorithm-based imbalanced learning strategies include clustering algorithms, deep learning, cost-sensitive learning, and extreme learning machine (ELM)96. The clustering algorithm could divide the samples in the space into different clusters, where the samples in the same cluster have similarities. After clustering the dataset, sampling the data according to representative samples such as cluster centers can effectively ensure the balance of the data size of different clusters. Deep learning uses the characteristics of algorithms to capture patterns in the imbalanced data to make classification and prediction more accuracy. Cost-sensitive learning guides the imbalanced learning process with the concept of 'cost'. The optimization goal of the algorithm is to minimize the total cost of classification errors by focusing on the samples with higher error costs. ELM, as the basic classifier of the ensemble network, can guarantee the accuracy of a single network with the combination of the ensemble methods to well improve the classification performance of imbalanced datasets. Lu et al.97 collected an imbalanced formability dataset of experimental HOIPs, including 539 HOIP and 24 non-HOIP samples. As shown in Fig. 7a, b, 9 different sampling methods including undersampling, oversampling and mixed sampling were introduced for unbalanced learning, while 10 different supervised and semi-supervised algorithms were used to select the best modeling algorithm to construct the model. After comparison, the mixed sampling method SMOTEENN has the best performance with the LOOCV accuracy and precision of the corresponding model have both reached 100%, and the accuracy of the test set has reached 95.5%. The LOOCV average accuracy of 100 random partitions and the average accuracy of the test set also exceeded 99.0%, respectively. The method of SHapley Additive exPlanations (SHAP) was used to extract and analyze important features of A-site atomic radius, A-site ionic radius, and tolerance factor to reveal the relationship with formability.

MACHINE LEARNING STRATEGIES FOR SMALL DATA

Machine learning strategies including active learning and transfer learning have been shown to be effective methods of handling small datasets in materials science.

Active learning

Active learning, also known as adaptive learning, is one of the key technologies for solving small data problems. The core of active learning is to select the samples from a large number of unlabeled data for labeling to make the information in the small data represent the large unlabeled data as much as possible to realize the analysis and processing of big data under small data98. The active learning workflow consists of the following steps: (1) train the model based on the labeled training set; (2) use the model to evaluate the acquisition function in the pool of unlabeled samples; (3) label the data points with the highest acquisition function scores; (4) add the labeled data points to the training set to train the model. The learning steps of active training, scoring, labeling, and acquisition are repeated until the model reaches sufficient accuracy. In the materials design based on active learning shown in Fig. 8a, the machine learning model would be constructed to design or screen out the candidate materials for further experimental or computational validation99. Then the verified candidate samples are taken back to the training set for modeling. Active learning can continuously enlarge the data size and improve the accuracy of the model in the process to realize the two-way optimization of data and model to be applied widely in materials machine learning with small data.

The core steps in the active learning workflow include the sampling, labeling, validation and evaluation of the significant samples from the unlabeled sample pool. The data sampling strategy used in the active learning process to filter out data points from the unlabeled sample pool is rather critical to improving the prediction accuracy of machine learning models. Common data sampling strategies include manual empirical sampling and Bayesian optimization sampling100. Manual empiri- cal sampling refers to the manual labeling of data samples by experts using expertize and traditional experience, which high- lights the importance of domain knowledge in machine learning. Bayesian optimization algorithms can automatically label the samples by using prior knowledge to approximate the posterior distribution of the unknown objective function. The basic idea of Bayesian optimization sampling is to balance the needs of 'exploration' and 'exploitation'. The 'exploitation' samples the most likely optimal solution region based on the posterior distribution; while the 'exploration' is usually to obtain sampling points in areas with low sampling density in order to improve the prediction accuracy of the model and reduce the fluctuation of prediction values101. The 'exploration' strategy is preferred in the initial stage when data size is insufficient, and it is more focused on improving the model prediction accuracy. As the data size gradually increases and the model prediction accuracy improves, the strategy gradually shifts to the 'exploitation' strategy, focusing on finding the optimal target value. The acquisition function is one of the cores of Bayesian optimization, which is used to evaluate and filter the most informative sample points from the unlabeled samples to be back to the original training set to perform active learning. Common acquisition functions include upper confidence bound (UCB), probability of improvement (PI) and expected improvement (EI), and Thompson Sampling101. In materials science, validation and evaluation of the selected labeled samples are usually performed through experiments or first-principles calculations. Active learning is an iteration process. Even if the model constructed with the original small data is not ideal, the size of modeling data and model accuracy can be improved through the

iteration of active learning. In addition, active learning also integrates machine learning well with experiments or first- principles calculations. The application of active learning in materials is no longer only in the theoretical stage, but combined with experiments or calculations through machine learning models to achieve the purpose of optimization.

In recent years, active learning has been widely applied in materials machine learning with small data. Xue et al.102 collected 22 Ni-Ti-based shape memory alloys and the thermal hysteresis property. The algorithm of SVR combined with efficient global optimization (EGO) search was applied to construct the thermal hysteresis prediction model to design Ni-Ti-based shape memory alloys with low thermal hysteresis. Models were trained multiple times and cross-validated with initial alloy data. After the model construction, EGO was used to search for 4 samples with low thermal hysteresis from the 800,000 searched spaces for experi- ments. After experimental validation, the 4 samples were put back into the training set for modeling-search-experiment iteration. Of the 36 samples searched after 9 iterations, 14 samples have thermal hysteresis smaller than any of the 22 samples in the original dataset, with Ti50.0Ni46.7Cu0.8Fe2.3Pd0.2 having the smallest thermal hysteresis of 1.84 K. Zhao et al.101 developed an effective active learning model to describe the relationship between elemental composition and hardness of 6061-aluminum alloy by combining high-throughput experiments and Bayesian optimized sampling strategy. First, 32 6061-aluminum alloys with different composition ratios were prepared and characterized for hardness using a full-flow high-throughput alloy preparation and character- ization system. 309 descriptors were constructed as initial features by elemental composition and alloy domain knowledge. After feature selection with variance, maximum information coefficient, weight coefficient, Pearson correlation coefficient, and sequence backward selection, the remaining 5 significant features were screened out for model construction. After comparing various algorithms, the SVR algorithm with kernel function of radial basis function was used to construct model to predict the hardness of aluminum alloys. Then, bootstrap was used to generate 1000 training datasets containing 32 samples by random sampling, and the above training datasets were used to obtain 1000 correspond- ing machine learning models for predicting the hardness of 33,600 candidates in the potential component space. Manual empirical sampling and Bayesian optimized sampling were used to select samples from the candidates for labeling and subsequent experiments, where the Bayesian sampling strategy specifically used 4 methods: the EGO algorithm, the knowledge gradient (KG) algorithm, the maximum hardness distribution method and the maximum error distribution method, each taking 4 data points and designing a total of 16 experimental alloy components for the next iteration of experiments at each step. The experimental data were returned to the initial dataset for further iterations of feature selection and model construction before convergence conditions were reached. After three iterations, the results showed that the adaptive sampling strategy of the Bayesian optimization algorithm could guide the experiments more effectively than manual empirical sampling, with a 63.03% reduction in MAE and a 53.85% reduction in RMSE. The hardness prediction RMSE of final model is 4.49 HV, which is close to the experimental error of 4.05 HV for the test sample. This work achieves the composition optimization of the hardness properties of 6061-aluminum alloy by the active learning strategy after Bayesian sampling optimiza- tion, which provides guidance for the design and performance optimization of other multi-alloy materials.

Transfer learning

Transfer learning refers to the acquisition of knowledge in a given source domain and learning task to help improve the learning of the predictive model in the target domain103. Transfer learning can be divided into model-based transfer learning, relation-based transfer learning and sample-based transfer learning according to transfer methods104. The model-based transfer learning method is to improve the prediction accuracy by adjusting the parameters of the pre-trained model. Relation-based transfer learning utilizes relations for analogical transfer such as cooking according to a recipe can be compared to conducting a scientific experiment according to a report. The sample-based transfer learning method is to directly assign different weights to different samples to complete the transfer. As shown in Fig. 8b, in the materials filed, transfer learning generally refers to model-based transfer learning by serving the small data in the target domain from the big data in the source domain105. After using the materials big data of the source domain to construct the pre-trained model, the parameters of the pre-trained model are adjusted in combination with the small data of the target domain to improve the prediction accuracy of the model to the small data.

Wu et al.106 developed a high-precision polymer thermal conductivity prediction model through transfer learning and Bayesian molecular design algorithm to screen out thousands of polymers with high thermal conductivity, of which 3 candidates were successfully synthesized and characterized after the experi- mental feasibility evaluation. A lot of polymer samples and the properties data were collected from the databases of PoLyInfo and QM9 to construct a pre-trained model107,108. After comparing different models, it was found that the pre-trained model of the heat capacity Cv owned the highest prediction accuracy. Then, the parameters of the pre-trained model were adjusted to be transferred to the prediction of thermal conductivity by the 28 samples with thermal conductivity. The results showed that the MAE of the model after transfer learning reached 0.0204 W (m · k)−1, which is 40% lower for directly trained models on each data point. Combined with Bayesian algorithm, a lot of repeating unit structures were designed for screening. Finally, 24 molecular structures were screened, of which 3 were successfully verified by experimental synthesis and characterization. The experimental results showed that the thermal conductivity of polymers assisted by transfer learning and Bayesian molecular design was higher than that of polymer materials in published papers. This research achievement also confirms that transfer learning and Bayesian molecular design can be successfully applied to the design and discovery of polymer materials. Lee et al.109 applied a crystal graph convolutional neural network (CGCNN) to a transfer learning model (TL-CGCNN) to improve the accuracy of material machine learning models with small data and quantitatively explored the effect of the sample size of the pre-trained and target models on the accuracy of the transfer learning model. The crystal structures and their corresponding bandgaps Eg and stratigraphic energies ΔEf were first collected from the Materials Project Database (MPD), a first-principles computational database. Then, three large datasets containing 10,000, 54,000, and 113,000 data, respectively,

were used to train the pre-trained models in conjunction with the CGCNN. In addition, bulk modulus (KVRH), dielectric constant (εr) and quasiparticle bandgap (GW-Eg) data were also collected to confirm the robustness of CGCNN for more cases with insufficient data volume. The prediction accuracy of Eg and ΔEf from pre- trained models with different sample sizes and comparison with conventional machine learning models reveals that the accuracy of TL-CGCNN models is much better than that of conventional machine learning models and the improvement of prediction ability is greater when the pre-trained models are trained with more data. The predictions of KVRH, εr and GW-Eg are also consistent with the above pattern and it is found that TL-CGCNN may be better for prediction model affected by small amount of data. The prediction of attributes in the target model by TL- CGCNN becomes more accurate when the pre-trained model is trained with larger data and the high correlation between the pre- trained model and the target model. Yamada et al.110 developed a pre-trained model library called XenonPy.MDL for transfer learning between different materials and their properties. The library has over 140,000 pre-trained models covering a wide range of materials including small molecules, polymers and inorganic crystalline materials. These pre-trained models are used to successfully span superior transferability between different materials and their properties, even beyond the different disciplines of materials science. This work provides a successful processing paradigm for small data materials machine learning using transfer learning and confirms the interconnectedness of almost all tasks in materials science, forging a bridge between small molecules and polymers, organic and inorganic materials. If the amount of collected target property data is very limited, but the amount of property data related to the target property is relatively abundant, transfer learning could be a very good choice.

CONCLUSION AND OUTLOOK

In this review, we discussed the dilemma of small data in materials machine learning and introduced the commonly used methods to deal with the small data machine learning from the aspects of data sources, algorithms, and machine learning strategies, including data extraction from publications, material database construction, high-throughput computations and experiments, small data modeling algorithms, imbalanced learning, active learning, and transfer learning. At present, the data size of most materials machine learning is still in the small data stage and will remain in the small data stage for a long time due to the inconsistency in the development progress, including the different types of materials, materials synthesis and characterization technology, materials classification and naming standards, data- base development technology, modeling algorithms and other factors. Therefore, handling small data modeling in materials machine learning is also one of the important directions. Here, we propose some future directions for further small data machine learning in materials science:

(1) Data management: In the past, data could be regarded as a series of apparent observations used to gain knowledge. But in the future, data would be more considered the information representing the results of complicated effects form multiple factors, which puts forward higher require- ments for data management111. Data management includes the processes of data collection, storage, screening, labeling, annotation, augmentation, evaluation, ablation and virtua- lization, which is a long-term process and requires the efforts of scholars and governments around the world. In addition, the dataset tends to be fixed and the machine learning is only mainly on the specific dataset in the previous materials machine learning process. But now, data iteration has been becoming the focus and more efforts are concentrated on improving the model performance through iterative models such as active learning, which also requires a more systematic approach to data management112.

(2) Methods combination: This review introduced a variety of small data machine learning methods and strategies in materials science. These methods and strategies should be used in combination to pursue better model performance, such as the combination of experiments and calculations, active learning, and materials database construction to develop a materials development system that integrates machine learning, database, experiments and computations. Besides, these methods should be more attached to the material design philosophy. Most applications of model in this review tend to screen materials or use patterns discovered by models to design materials, which is called forward design. In the future, inverse design based on small datasets to deduce the composition and structure of materials according to the required properties is also one of the future development directions.

(3) Machine learning algorithms and strategies: In addition to materials data, machine learning algorithms and strategies are also important factors to determine the applications. This review introduced several modeling algorithms suitable for small datasets, but different algorithms are only suitable for the specific data, and there is no modeling algorithm that is generally applicable to small data. Therefore, more modeling algorithms suitable for small data still need to be developed. Similarly, machine learning strategies suitable for small data, including active learning and transfer learning, also have great potential for application and development.

This development should take into account both data and algorithms, not only to establish a complete database or use the existing technologies to integrate materials data to increase the data size, but also to continuously develop small data modeling algorithms and machine learning strategies. In the future, machine learning will continue to occupy an increasingly important place in materials design and discovery. Especially for experimenters, this review would help better handle the precious and limited experimental data with machine learning methods to accelerate material design and discovery. A drop of water can refract the brilliance of the sun. Similarly, small materials data can also be used to explore mysterious and interesting patterns in the vast world of materials through machine learning.

DATA AVAILABILITY

All the data of the cases could be obtained from the corresponding references.

CODE AVAILABILITY

All the code of the cases could be obtained from the corresponding references.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 52102140), Shanghai Pujiang Program (No.21PJD024), and the Key Research Project of Zhejiang Labora-

tory (No. 2021PE0AC02).

AUTHOR CONTRIBUTIONS

P.X. collected publications and completed the framework of the manuscript. X.J., M.L., and W.L. revised the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01000-z.

REFERENCES

1. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Electron. Mark. 31, 685–695 (2021).
2. Bi, Q., Goodman, K. E., Kaminsky, J. & Lessler, J. What is machine learning? A primer for the epidemiologist. Am. J. Epidemiol. 188, 2222–2239 (2019).
3. Warin, T. & Stojkov, A. Machine learning in finance: a metadata-based systematic review of the literature. J. Risk Financ. Manag. 14, 302 (2021).
4. Ahmed, S., Alshater, M. M., Ammari, A. E. & Hammami, H. Artificial intelligence and machine learning in finance: A bibliometric review. Res. Int. Bus. Financ. 61, 101646 (2022).
5. Mueller, B., Kinoshita, T., Peebles, A., Graber, M. A. & Lee, S. Artificial intelligence and machine learning in emergency medicine: a narrative review. Acute. Med. Surg. 9, e740 (2022).
6. Sabry, F., Eltaras, T., Labda, W., Alzoubi, K. & Malluhi, Q. Machine learning for healthcare wearable devices: the big picture. J. Healthc. Eng. 2022, 4653923 (2022).
7. Okoroafor, E. R. et al. Machine learning in subsurface geothermal energy: two decades in review. Geothermics 102, 102401 (2022).
8. Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A. & De Felice, F. Artificial Intelli- gence and machine learning applications in smart production: progress, trends, and directions. Sustainability 12, 492 (2020).
9. Crampon, K., Giorkallos, A., Deldossi, M., Baud, S. & Steffenel, L. A. Machine- learning methods for ligand-protein molecular docking. Drug Discov. Today 27, 151–164 (2022).
10. Jiang, Y., Luo, J., Huang, D., Liu, Y. & Li, D. D. Machine learning advances in microbiology: a review of methods and applications. Front. Microbiol. 13, 925454 (2022).
11. Cai, J., Chu, X., Xu, K., Li, H. & Wei, J. Machine learning-driven new material discovery. Nanoscale Adv. 2, 3115–3130 (2020).
12. Chan, C. H., Sun, M. & Huang, B. Application of machine learning for advanced material prediction and design. Eco. Mat. 4, e12194 (2022).
13. Zhu, L., Zhou, J. & Sun, Z. Materials data toward machine learning: advances and challenges. J. Phys. Chem. Lett. 13, 3965–3977 (2022).
14. Yang, C. et al. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. Acta Mater. 222, 117431 (2022).
15. Tao, Q., Xu, P., Li, M. & Lu, W. Machine learning for perovskite materials design and discovery. npj Comput. Mater. 7, 23 (2021).
16. Liu, X. et al. Material machine learning for alloys: applications, challenges and perspectives. J. Alloy. Compd. 921, 165984 (2022).
17. Xu, P., Chen, H., Li, M. & Lu, W. New opportunity: machine learning for polymer materials design and discovery. Adv. Theor. Simul. 5, 2100565 (2022).
18. Faraway, J. J. & Augustin, N. H. When small data beats big data. Stat. Probabil. Lett. 136, 142–145 (2018).
19. Chandrasekaran, V. & Jordan, M. I. Computational and statistical tradeoffs via convex relaxation. Proc. Natl Acad. Sci. USA 110, E1181–E1190 (2013).
20. Zhang, Q., Chang, D., Zhai, X. & Lu, W. OCPMDM: Online computation platform for materials data mining. Chemom. Intell. Lab. 177, 26–34 (2018).
21. Li, L. et al. Studies on the regularity of perovskite formation via machine learning. Comput. Mater. Sci. 199, 110712 (2021).
22. Yang, X., Li, L., Tao, Q., Lu, W. & Li, M. Rapid discovery of narrow bandgap oxide double perovskites using machine learning. Comput. Mater. Sci. 196, 110528 (2021).
23. Tao, Q. et al. Multiobjective stepwise design strategy-assisted design of high- performance perovskite oxide photocatalysts. J. Phys. Chem. C. 125, 21141–21150 (2021).
24. Xu, P. et al. Search for ABO3 type ferroelectric perovskites with targeted multi- properties by machine learning strategies. J. Chem. Inf. Model. 62, 5038–5049 (2022).
25. Schwarz, K. & Sundararaman, R. The electrochemical interface in first-principles calculations. Surf. Sci. Rep. 75, 100492 (2020).
26. Liu, B. et al. Application of high-throughput first-principles calculations in ceramic innovation. J. Mater. Sci. Technol. 88, 143–157 (2021).
27. Dardzinski, D., Yu, M., Moayedpour, S. & Marom, N. Best practices for first- principles simulations of epitaxial inorganic interfaces. J. Phys. Condens. Matter 34, 233002 (2022).
28. Fjodorova, N. & Novic, M. Integration of QSAR and SAR methods for the mechanistic interpretation of predictive models for carcinogenicity. Comput. Struct. Biotechnol. J. 1, e201207003 (2012).
29. Moussaoui, M., Laidi, M., Hanini, S. & Hentabli, M. Artificial neural network and support vector regression applied in quantitative structure-property relationship modelling of solubility of solid solutes in supercritical CO2. Kem. u. industriji. 69, 611–630 (2020).
30. Zhang, K. & Zhang, H. Predicting solute descriptors for organic chemicals by a deep neural network (dnn) using basic chemical structures and a surrogate metric. Environ. Sci. Technol. 56, 2054–2064 (2022).
31. Beckner, W., Mao, C. M. & Pfaendtner, J. Statistical models are able to predict ionic liquid viscosity across a wide range of chemical functionalities and experimental conditions.

Mol. Syst. Des. Eng. 3, 253–263 (2018).

32. Lian, Z., Li, M. & Lu, W. Fatigue life prediction of aluminum alloy via knowledge- based machine learning. Int. J. Fatigue 157, 106716 (2022).

33. Li, Y., Li, T. & Liu, H. Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53, 551–577 (2017).

34. Khaire, U. M. & Dhanalakshmi, R. Stability of feature selection algorithm: a review. J. King Saud. Univ. Com. 34, 1060–1073 (2022).

35. France, S. L. & Akkucuk, U. A review, framework, and R toolkit for exploring, evaluating, and comparing visualization methods. Vis. Comput 37, 457–475 (2020).

36. Jia, W., Sun, M., Lian, J. & Hou, S. Feature dimensionality reduction: a review. Complex Intell. Syst. 8, 2663–2693 (2022).

37. Xie, Y. & Sun, P. Terahertz data combined with principal component analysis applied for visual classification of materials. Opt. Quant. Electron. 50, 46 (2018).

38. Tula, T. et al. Machine learning approach to muon spectroscopy analysis. J. Phys. Condens. Matter 33, 194002 (2021).

39. Gardner-Lubbe, S. Linear discriminant analysis for multiple functional data analysis. J. Appl. Stat. 48, 1917–1933 (2021).

40. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descrip- tor in an immensity of offered candidates. Phys. Rev. Mater. 2, 083802 (2018).

41. Ouyang, R., Ahmetcik, E., Carbogno, C., Scheffler, M. & Ghiringhelli, L. M. Simultaneous learning of several materials properties from incomplete data- bases with multi-task SISSO. J. Phys. Mater. 2, 024002 (2019).

42. He, J. & Fan, X. Evaluating the performance of the k-fold cross-validation approach for model selection in growth mixture modeling. Struct. Equ. Model. 26, 66–79 (2018).

43. Zhang, J. & Wang, S. A fast leave-one-out cross-validation for SVM-like family. Neural Comput. Appl. 27, 1717–1730 (2015).

44. Lu, K. et al. Machine learning model for high-throughput screening of perovskite manganites with the highest néel temperature. J. Supercond. Nov. Magn. 34, 1961–1969 (2021).

45. Erickson, M. E., Ngongang, M. & Rasulev, B. A refractive index study of a diverse set of polymeric materials by QSPR with quantum-chemical and additive descriptors. Molecules 25, 3772 (2020).

46. Shibayama, S. & Funatsu, K. Investigation of preprocessing and validation methodologies for PAT: case study of the granulation and coating steps for the manufacturing of ethenzamide tablets. AAPS Pharm. Sci. Tech. 22, 41 (2021).

47. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic mate- rials synthesis parameters with deep learning. npj Comput. Mater. 3, 2096–5001 (2017).

48. Kajita, S., Ohba, N., Suzumura, A., Tajima, S. & Asahi, R. Discovery of superionic conductors by ensemble-scope descriptor. NPG Asia Mater. 12, 31 (2020).

49. Tao, Q. et al. Machine learning aided design of perovskite oxide materials for photocatalytic water splitting. J. Energy Chem. 60, 351–359 (2021).

50. Zeng, Z., Shi, H., Wu, Y. & Hong, Z. Survey of natural language processing techniques in bioinformatics. Comput. Math. Methods Med. 2015, 674296 (2015).

51. Perovšek, M., Kranjc, J., Erjavec, T., Cestnik, B. & Lavrač, N. TextFlows: a visual programming platform for text mining and natural language processing. Sci. Comput. Program. 121, 128–152 (2016).

52. Kononova, O. et al. Opportunities and challenges of text mining in aterials research. iScience 24, 102155 (2021).

53. Hong, Z., Ward, L., Chard, K., Blaiszik, B. & Foster, I. Challenges and advances in information extraction from scientific literature: a review. JOM 73, 3383–3400 (2021).

54. Memon, J., Sami, M., Khan, R. A. & Uddin, M. Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access 8, 142642–142668 (2020).

55. Dalva, D., Guz, U. & Gurkan, H. Effective semi-supervised learning strategies for automatic sentence segmentation. Pattern Recogn. Lett. 105, 76–86 (2018).

56. Leaman, R., Wei, C. H. & Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J. Cheminforma. 7, S3 (2015).

57. Maksutov, A. A., Zamyatovskiy, V. I., Morozov, V. O. & Dmitriev, S. O. The Transformer Neural Network Architecture for Part-of-Speech Tagging. ElConRus 536–540 (IEEE, 2021).

58. Phillips, S. L. C. Aligning grammatical theories and language processing models. J. Psycholinguist. Res. 44, 27–46 (2015).

59. Lewis, D. D. & Jones, K. S. Natural language processing for information retrieval. Commun. ACM 39, 92–101 (1996).

60. Goyal, A., Gupta, V. & Kumar, M. Recent named entity recognition and classifi- cation techniques: a systematic review. Comput. Sci. Rev. 29, 21–43 (2018).

61. Safaa Eltyeb, N. S. Chemical named entities recognition: a review on approaches and applications. J. Cheminformatics 6, 1–12 (2014).

62. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J. Chem. Inf. Model. 56, 1894–1904 (2016).

63. Mavracic, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. ChemDataExtractor 2.0: autopopulated ontologies for materials science. J. Chem. Inf. Model. 61, 4280–4289 (2021).

64. Katsura, Y. et al. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Sci. Technol. Adv. Mat. 20, 511–520 (2019).

65. Wang, Z. et al. Dataset of solution-based inorganic materials synthesis proce- dures extracted from the scientific literature. Sci. Data 9, 231 (2022).

66. Yin, H.-Q. et al. The materials data ecosystem: materials data science and its role in data-driven materials discovery. Chin. Phys. B 27, 118101 (2018).

67. Xu, Y. Accomplishment and challenge of materials database toward big data. Chin. Phys. B 27, 118901 (2018).

68. Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. ACS Macro. Lett. 6, 1078–1082 (2017).

69. Zixin, L. et al. Materials science database in material research and development: recent applications and prospects. Front. Data Comput. 2020, 78–90 (2020).

70. Huang, Y., Shang, M., Liu, T. & Wang, K. High-throughput methods for genome editing: the more the better. Plant Physiol. 188, 1731–1745 (2022).

71. He, X., Zhang, N., Cao, W., Xing, Y. & Yang, N. Application progress of high- throughput sequencing in ocular diseases. J. Clin. Med. 11, 3485 (2022).

72. Xiaoli, F. Materials genome initiative and first-principles high-throughput com- putation. Mater. China 34, 689–695 (2015).

73. Curtarolo, S. et al. The high-throughput highway to computational materials design. Nat. Mater. 12, 191–201 (2013).

74. Shulin, L., Tianshu, L., Xinjiang, W., Muhammad, F. & Lijun, Z. High-throughput computational materials screening and discovery of optoelectronic semi- conductors. WIREs Comput. Mol. Sci. 11, e1489 (2021).

75. Liu, Y. et al. High-throughput experiments facilitate materials innovation: a review. Sci. China Technol. Sc. 62, 521–545 (2019).

76. Hu, W., Zhang, L. & Pan, Z. Designing two-dimensional halide perovskites based on high-throughput calculations and machine learning. ACS Appl. Mater. Inter- faces 14, 21596–21604 (2022).

77. Hayashi, Y., Shiomi, J., Morikawa, J. & Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. npj Comput. Mater. 8, 222 (2022).

78. Zhao, Y. et al. Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. Nat. Commun. 12, 2191 (2021).

79. Kirkpatrick, J. et al. Pushing the frontiers of density functionals by solving the fractional electron problem. Science 374, 1385–1389 (2021).

80. Takamoto, S. et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. Nat. Commun. 13, 2991 (2022).

81. Scholkopf, A. J. S. B. A tutorial on support vector regression. Stat. Comput. 14, 199–222 (2004).

82. Shawe-Taylor, J. & Sun, S. A review of optimization methodologies in support vector machines. Neurocomputing 74, 3609–3618 (2011).

83. Deringer, V. L. et al. Gaussian process regression for materials and molecules. Chem. Rev. 121, 10073–10141 (2021).

84. Talekar, B. A detailed review on decision tree and random forest. Biosci. Biotech. Res. C. 13, 245–248 (2020).

85. Biau, G., Cadre, B. & Rouvière, L. Accelerated gradient boosting. Mach. Learn. 108, 971–992 (2019).

86. Duan, J., Asteris, P. G., Nguyen, H., Bui, X.-N. & Moayedi, H. A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ICA-XGBoost model. Eng. Comput. 37, 3329–3346 (2020).

87. Afzal, W. & Torkar, R. On the application of genetic programming for software engineering predictive modeling: A systematic review. Expert Syst. Appl. 38, 11984–11997 (2011).

88. Guo, Z., Hu, S., Han, Z. K. & Ouyang, R. Improving symbolic regression for pre- dicting materials properties with iterative variable selection. J. Chem. Theory Comput. 18, 4945–4951 (2022).

89. Weng, B. et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. Nat. Commun. 11, 3513 (2020).

90. Shi, L., Chang, D., Ji, X. & Lu, W. Using data mining to search for perovskite materials with higher specific surface area. J. Chem. Inf. Model. 58, 2420–2427 (2018).

91. Lu, K., Chang, D., Ji, X., Li, M. & Lu, W. Machine learning aided discovery of the layered double hydroxides with the largest basal spacing for super-capacitors. Int. J. Electrochem. Sc. 16, 211146 (2021).

92. Chen, H., Shang, Z., Lu, W., Li, M. & Tan, F. A property□ driven stepwise design strategy for multiple low□ melting alloys via machine learning. Adv. Eng. Mater. 23, 2100612 (2021).

93. Lu, T., Li, H., Li, M., Wang, S. & Lu, W. Inverse design of hybrid organic-inorganic perovskites with suitable bandgaps via proactive searching progress. ACS Omega 7, 21583–21594 (2022).

94. Haibo, H. & Garcia, E. A. Learning from imbalanced data. IEEE T. Knowl. Data En. 21, 1263–1284 (2009).

95. Li, Y.-X., Chai, Y., Hu, Y.-Q. & Yin, H.-P. Review of imbalanced data classification methods. Control Decis. 34, 673–688 (2019).

96. Wang, L., Han, M., Li, X., Zhang, N. & Cheng, H. Review of classification methods for unbalanced data sets. Comput. Eng. Appl. 57, 42–52 (2021).

97. Lu, T., Li, H., Li, M., Wang, S. & Lu, W. Predicting experimental formability of hybrid organic-inorganic perovskites via imbalanced learning. J. Phys. Chem. Lett. 13, 3032–3038 (2022).

98. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput. Mater. 5, 21 (2019).

99. Xin, R. et al. Active-learning-based generative design for the discovery of wide- band-gap materials. J. Phys. Chem. C. 125, 16118–16128 (2021).

100. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat. Commun. 11, 5966 (2020).

101. Zhao, W. et al. Composition refinement of 6061 aluminum alloy using active machine learning model based on bayesian optimization sampling. Acta Metall. Sin. 57, 797–809 (2021).

102. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. Nat. Commun. 7, 11241 (2016).

103. Ranaweera, M. & Mahmoud, Q. H. Virtual to real-world transfer learning: a sys- tematic review. Electronics 10, 1491 (2021).

104. Zhuang, F. et al. A comprehensive survey on transfer learning. P. IEEE 109, 43–76. (2021).

105. Schutt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A. & Muller, K. R. SchNet—A deep learning architecture for molecules and materials. J. Chem. Phys. 148, 241722 (2018).

106. Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput. Mater. 5, 5 (2019).

107. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset's che- mical diversity limits the generalizability of machine learning predictions. J. Cheminforma. 11, 69 (2019).

108. Ma, R. & Luo, T. PI1M: a benchmark database for polymer informatics. J. Chem. Inf. Model. 60, 4684–4690 (2020).

109. Lee, J. & Asahi, R. Transfer learning for materials informatics using crystal graph convolutional neural network. Comput. Mater. Sci. 190, 110314 (2021).

110. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. ACS Cent. Sci. 5, 1717–1730 (2019).

111. Hong, W., Xiang, X.-D. & Lanting, Z. On the data-driven materials innovation infrastructure. Engineering 6, 609–611 (2020).

112. Weixin, L. et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat. Mach. Intell. 4, 904 (2022). 21