Xu *et al. J Mater Inf* 2023;3:16
DOI: 10.20517/jmi.2023.18

Journal of Materials Informatics

Review

Open Access

Check for updates

Virtual sample generation in machine learning assisted materials design and discovery

Pengcheng Xu[1], Xiaobo Ji[2], Minjie Li[2,*], Wencong Lu[2,3,4,*]

[1]Materials Genome Institute, Shanghai University, Shanghai 200444, China.
[2]Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China.
[3]Zhejiang Laboratory, Hangzhou 311100, Zhejiang, China.
[4]Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education, Shanghai 200444, China.

[*]**Correspondence to:** Prof. Minjie Li, Prof. Wencong Lu, Department of Chemistry, College of Sciences, Shanghai University, 99 Shangda Road Baoshan District, Shanghai 200444, China. E-mail: minjieli@shu.edu.cn; wclu@shu.edu.cn

How to cite this article: Xu P, Ji X, Li M, Lu W. Virtual sample generation in machine learning assisted materials design and discovery. *J Mater Inf* 2023;3:16. https://dx.doi.org/10.20517/jmi.2023.18

Received: 2 May 2023 **First Decision:** 6 Jun 2023 **Revised:** 24 Jun 2023 **Accepted:** 29 Jun 2023 **Published:** 13 Jul 2023

Academic Editor: Xingjun Liu **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

Abstract

Virtual sample generation (VSG), as a cutting-edge technique, has been successfully applied in machine learning-assisted materials design and discovery. A virtual sample without experimental validation is defined as an unknown sample, which is either expanded from the original data distribution for modeling or designed via algorithms for predicting. This review aims to discuss the applications of VSG techniques in machine learning-assisted materials design and discovery based on the research progress in recent years. First, we summarize the commonly used VSG algorithms in materials design and discovery for data expansion of the training set, including Bootstrap, Monte Carlo, particle swarm optimization, mega trend diffusion, Gaussian mixture model, random forest, and generative adversarial networks. Next, frequently employed searching algorithms for materials discovery are introduced, including particle swarm optimization, efficient global optimization, and proactive searching progress. Then, universally adopted inverse design methods are presented, including genetic algorithm, Bayesian optimization, and pattern recognition inverse projection. Finally, the future directions of VSG in the design and discovery of materials are proposed.

Keywords: Materials machine learning, virtual sample generation, searching algorithms, inverse design

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

INTRODUCTION

Machine learning, as a branch of artificial intelligence, is an interdisciplinary field that encompasses statistics, computer science, and engineering. It enables automatic learning, pattern recognition, and decision-making based on data analysis to facilitate the discovery of hidden patterns and regularities[1-5]. In materials science, machine learning has become widely applied to accelerate the design and discovery of new materials in recent decades[6-9]. For instance, machine learning can construct models to predict the physical properties, chemical reactivity, and structural stability of new materials[10-14]. By combining experimental data with advanced machine learning algorithms, researchers can develop a powerful predictive and generalizable model to optimize materials design and reduce the trial-and-error costs. Furthermore, machine learning can aid in materials structure classification and prediction and the optimization of materials preparation and processing. Currently, machine learning-assisted acceleration of materials design and discovery has been widely applied to various materials systems, including alloys, perovskites, organic molecules, and polymers[15-19].

However, the development of new materials requires the characterization of the properties and structures through experiments. Due to the high cost of acquiring experimental data, even with the first-principle calculations, a significant amount of resources can be consumed for the complex material systems. As a result, the small number of samples would make it difficult to construct accurate models with ideal performance, leading to the small sample dilemma in materials machine learning[20]. Moreover, samples prepared through experiments tend to have uncertain and unstable performance. The quality and stability of data could be influenced by various factors of preparation processes and testing conditions, which makes data processing rather challenging as the same materials may produce different data results under different experimental conditions. Additionally, searching for target materials from the vast materials space, whether through experiments or first-principle calculations, is a significant burden for researchers.

Many researchers have developed various methods to tackle the small sample dilemma in materials machine learning, including data extraction from publications, materials database construction, high-throughput experiments and computations, transfer learning, and active learning[21-25]. However, these methods usually have limited effectiveness for a given small dataset. Nevertheless, after collecting a small dataset, virtual sample generation (VSG) technology can be used to increase the data size. A virtual sample without experimental validation is defined as an unknown sample, which is either expanded from the original data distribution for modeling or designed via algorithms for predicting. VSG technology can generate effective virtual samples by utilizing the distribution information or prior knowledge of the original data to improve the accuracy and generalization ability of the model. In addition, VSG technology can also be used in materials machine learning to screen out the target materials in a large number of virtual samples generated by machine learning models for prediction, search for target materials in the materials space, and inverse design the target materials.

This review aims to discuss the applications of VSG technology in materials, covering its progress, prospects, challenges, and controversies. Firstly, we introduce VSG methods to be applied to expand the training data size based on statistical analysis and modeling algorithms. Secondly, we summarize searching algorithms that have been successfully applied in materials science. Then, we introduce materials inverse design methods. Finally, we propose several development directions for VSG technology in materials science.

STATISTICS AND MODELING ALGORITHMS BASED VSG

In this section, we will introduce the commonly used statistics and modeling algorithm-based VSG methods for data expansion, including Bootstrap, Monte Carlo, particle swarm optimization (PSO) algorithm, mega trend diffusion (MTD), Gaussian mixture model (GMM), random forest (RF), and generative adversarial networks (GANs).

Bootstrap

Bootstrap is a resampling-based technique in statistical learning to estimate standard errors, confidence intervals, and biases[26]. As shown in Figure 1, the core of Bootstrap is to calculate the confidence interval of an estimator by repeated sampling from the original data. A new dataset formed from the original data by repeated sampling with replacement is called Bootstrap data, with the true values of the unknown information estimated by calculating the statistical parameters of these Bootstrap samples[27,28]. In Bootstrap, there exist columns of independently and identically distributed samples $X = [x_1, x_2, \ldots x_m]$ with the distribution function of F_m. By sampling $X^* = [x_1^*, x_2^*, \ldots, x_m^*]$ from X, and the distribution function F_m of the X^* could be obtained. If $\hat{F}_m$ is a good estimate of F_m, then the relationship between X and $\hat{F}_m$ can be fully reflected in the relationship between X^* and $\hat{F}_m$, and the statistical distribution of the overall samples can be estimated by repeated sampling. The Bootstrap method can be used to generate virtual samples when the data size cannot meet the modeling needs. The detailed steps of using Bootstrap to generate virtual samples from small original data are described as follows:

Arrange the original dataset $X = [x_1, x_2, \ldots x_m]$ in ascending order to obtain ascending statistics and randomly generate integers $i_1, i_2, \ldots i_m \in [1, m]$.

According to the subscripts corresponding to the generated integers $i_1, i_2, \ldots i_m \in [1, m]$, perform the repeated sampling with replacement on the original small-sample dataset $X = [x_1, x_2, \ldots x_m]$ to obtain a new Bootstrap dataset $X^* = [x_{i1}, x_{i2}, \ldots, x_{im}]$.

Repeat step (2) n times to obtain a virtual dataset $X^* = [x_1^*, x_2^*, \ldots, x_m^*]$ that can reflect the overall characteristics, and the generated virtual sample data volume is m^*n.

The advantage of Bootstrap lies in the fact that it does not require the assumptions about the distribution of samples in advance. In general, accurately determining the distribution of samples can be challenging. When analyzing the distribution of a dataset, if the assumptions made deviate significantly from the true distribution, substantial errors may occur. Furthermore, results obtained based on assumptions about the sample distribution often necessitate hypothesis testing to assess the reasonableness of those assumptions. Bootstrap is a non-parametric method by repeated sampling only with replacement, where the population estimators can be derived from sample estimators. In addition, Bootstrap has obvious advantages over other statistical methods under the conditions of complex or unknown data distribution. As long as there exist the original samples, plenty of virtual samples could be generated by simply repeated sampling with Bootstrap for modeling. However, it is also important to acknowledge the limitations of the Bootstrap method. Since Bootstrap relies heavily on repeated sampling, the accuracy and reliability of the virtual samples can be significantly compromised when the dataset is extremely small or not representative of the overall population. Moreover, Bootstrap could not analyze and extract potentially valuable information from the samples. Consequently, the generated virtual samples may not effectively bridge the information gap arising from the limited size of the original sample set. This limitation may lead to significant errors when employing the model to predict unknown datasets beyond the scope of the original data. Furthermore, the characteristic of random repeated sampling could result in substantial variability in the distribution of the

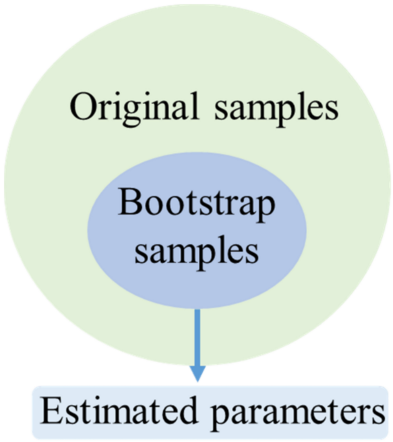

Figure 1. The scheme of Bootstrap.

generated samples. Particularly when dealing with sparse sample sizes, the probability density distribution of the virtual samples may deviate significantly from the assumed Gaussian distribution. This deviation can affect the accuracy and reliability of statistical analysis and predictions based on the generated samples. Rubin *et al.* proposed an enhanced version of the Bootstrap method named Bayesian Bootstrap, which combines Bootstrap with Bayesian algorithms to sample the original data[29]. Different from the traditional Bootstrap approach, which randomly samples with replacement from the data, Bayesian Bootstrap treats the weight vector in the empirical distribution function as an unknown parameter and simulates its posterior distribution. This integration of Bayesian principles enables a more refined sampling process. Notably, the samples generated by Bayesian Bootstrap tend to exhibit a probability density distribution that closely approximates a Gaussian distribution. This improvement enhances the reliability and accuracy of statistical analyses and predictions based on the generated samples.

The VSG method based on Bootstrap has been applied to materials machine learning with small data. Zhu *et al.* used Bootstrap-VSG combined with an extreme learning machine (ELM) to construct a prediction model of acetic acid consumption in a purified terephthalic acid (PTA) process[27]. The results show that compared with the original 30 samples, the ELM model constructed with 90 samples improves the prediction accuracy of the model with 30 samples by 27.48% after Bootstrap generates 60 virtual samples. Based on Bayesian Bootstrap, Han *et al.* proposed a Bootstrap-Bayesian dynamic modification model for the small sample size[28]. Firstly, Bootstrap confidence interval estimation was used to obtain the interval range at the mean 95% confidence level and generate Bootstrap samples that are uniformly distributed over the interval. Secondly, Bayesian parameter estimation was used to achieve Bayesian posterior estimates of the Bootstrap-based mean as virtual samples to extend the training sample set. The advantage of the Bootstrap-Bayesian dynamic modification model lies in that the virtual samples generated based on Bootstrap confidence intervals can provide more valid sample information. The Bootstrap-Bayesian dynamic modification model has been successfully applied to the underwater target classification, but few relevant applications have been found in the materials field, indicating the great potential for applications in materials research.

Monte Carlo

The basic principle of Monte Carlo is to use random sampling techniques to approximate the simulation probability, obtaining the simulation results on the basis of random repeated sampling and statistical analysis[30]. In terms of calculating the probability of an event occurrence, the Monte Carlo method can

estimate the corresponding probability by repeated sampling and calculating the frequency of the event. Monte Carlo is based on a probability model. Different from Bootstrap, which samples from the original sample, Monte Carlo analyzes the data distribution of the samples and then conducts random repeated sampling according to the data distribution. Consequently, the virtual samples generated by Monte Carlo align more closely with the probability distribution of the original data. According to the process described by the probability model, the results are simulated as an approximate solution to the problem, which includes three main steps:

The construction of the probability distribution model: Select an appropriate probability distribution to create a probability model to convert a non-random problem into a random problem.

Sampling according to the probability distribution: The core of the Monte Carlo method is sampling, which can make the obtained random numbers conform to the constructed probability distribution model to obtain samples that satisfy a certain probability distribution. Commonly used Monte Carlo sampling methods include rejection sampling, inverse sampling, acceptance-rejection sampling, importance sampling, Markov Monte Carlo sampling, and Latin Hypercube sampling.

The calculation of the estimator: After obtaining the samples that conform to the probability distribution model, it is required to calculate the statistical parameters of each sample as the estimator of the overall samples to investigate the simulation results and obtain the solution to the problem.

The VSG could be achieved with the Monte Carlo method by treating each input variable as a random variable, constructing the probability model based on the statistical characteristics[31]. Then the virtual samples reflecting the spatial characteristics of the probability distribution could be generated by sampling. The advantage of Monte Carlo lies in the principle of only repeated sampling, leading to a simpler calculation method and procedure. For stochastic tasks, Monte Carlo methods excel in the ability to directly simulate solutions that closely approximate the real solution. However, Monte Carlo relies on random sampling, while computers can only generate pseudo-random numbers instead of true random numbers. Moreover, unlike traditional methods that can precisely estimate the error, Monte Carlo is based on a probabilistic model that can only guarantee the error meets the accuracy requirement with a certain probability. The sampling technique in Monte Carlo is completely random, indicating that the samples can be located anywhere within the range of the input data distribution. Samples are more likely to be drawn from regions of the distribution with a high probability of occurrence. However, if the number of samples is small, samples with low probability may not be drawn in sufficient numbers. As a result, the samples tend to cluster in areas with a high probability of occurrence and may not accurately represent the distribution of the original sample.

Particle swarm optimization

PSO originates from the study of predation behavior of birds to seek the optimal solution through cooperation and information sharing among individuals in the group[32]. It is assumed that a flock of birds is randomly searching for a piece of food in an area. All the birds do not know the exact location of the food, but they know how far the current location is from the food. The flocks let other birds know their location by passing the respective information to each other during the searching process to figure out whether they have found the optimal solution for the food location. At the same time, the information of the optimal solution could also be transmitted to the whole group. Eventually, the entire flock of birds can gather around the food to find the optimal solution. In PSO, every solution to the optimization problem is a bird in the search space, which could also be called a “particle”. All the particles have the respective fitness value

determined by the optimized function and a velocity to determine the searching direction and distance. PSO is initialized as a group of random particles to iterate to find the optimal solution. In the iteration, each particle is characterized by its position and velocity, which use their individual historical optimal positions (p_{best}) and the group historical optimal positions of all other members (g_{best}) to find the searching direction and update their positions and velocities until the convergence conditions are met. The virtual sample generated by PSO is to obtain the optimal virtual sample by searching in the sample space. The particle swarm algorithm aims to minimize the error between the actual value and the expected value in the generation of virtual samples[33-34]. In order to reduce the error between the actual value and the predicted value, PSO searches for a better combination in the input data to ensure the validity of the virtual samples, where the process is transformed into a nonlinear constrained optimization problem:

$$\begin{aligned} &\text{Minimize } f(\mathrm{x}) \\ &\text{Subject to } x_{iL} \leqslant x_i \leq x_{iU};\ i = 1, 2 \ldots a \\ &G_k(x) \leqslant 0;\ k = 1, 2 \ldots c \end{aligned}$$

Where $f(\mathrm{x})$ is the objective function; x_{iL} and x_{iU} are the lower and upper limits of the search area of the input variable x; x_i is the input variable; a is the number of input attribute variables; c is the number of constraints; $G_k(x)$ is a nonlinear constraint. The solutions to the problem obtained by PSO are the optimal virtual samples. The advantages of PSO include its minimal number of adjustable parameters, the ability to generate virtual samples that closely approximate real values, and the absence of a requirement to know the distribution of samples in advance. However, there are also limitations to consider. Due to the limited number of adjustable parameters, the selection of parameter settings can greatly affect the model performance. Furthermore, the particle swarm tends to exhibit "convergence" as it approaches the optimal solution, causing the particle speeds to decrease and potentially resulting in a local optimal solution. Currently, PSO has been successfully applied to multi-objective optimization tasks, and this PSO-based VSG technique offers valuable insights for the multi-objective optimization of materials.

Gong *et al.* proposed the VSG technique of MC-PSO, which efficiently combines Monte Carlo and PSO algorithms[31]. Firstly, the ELM is combined with the original training set of small data to construct a model for prediction. Then the probability distribution of the original dataset is evaluated and sampled using the Monte Carlo method, with the sampled data used as the initial points for the PSO searching. The feature values of the virtual samples could be obtained by PSO searching. The process of PSO searching is terminated when the expected value of the fitness function or the maximum number of iterations is satisfied. The target variable of the virtual samples is calculated by the constructed ELM model. Finally, the virtual samples would be added back to the original dataset to train the new ELM model to evaluate the performance with the test set. MC-PSO has been successfully applied to the ethylene industry capacity prediction. Compared with the initial 30 samples, the root mean square error (RMSE) of the ELM model decreased from 0.0387 to 0.0220 after 60 virtual samples were generated and added to the dataset by MC-PSO.

Mega trend diffusion

The theoretical basis of the MTD technology is diffusion neural networks, which combine information diffusion with neural networks and regard data points as a data center with a fuzzy normal distribution in a certain interval[35]. Two new sample points are spread symmetrically on both sides of these data points using a symmetric spread function. Therefore, each sample can obtain two virtual samples after the diffusion to expand the data size of the original small data and fill the information gap caused by small data. Li *et al.*

took the integrity of the sample data distribution into consideration and proposed the method of MTD based on diffusion neural networks[36]. Different from diffusion neural networks, MTD considers the integrity of the data and generalizes the single-point diffusion to the overall diffusion, which could asymmetrically expand the attribute domain of the sample with the possibility of the occurrence of the sample reflected by the triangular membership function value[37]. The scheme of MTD is shown in Figure 2, and the steps to generate virtual samples with MTD technology are as follows:

In the dataset $X = \{x_1, x_2, \ldots x_n\}$, *max* and *min* represent the maximum and minimum values of a certain property of the sample; *CL* is the center point of the sample property to be calculated by the formula of (*max*+*min*)/2; N_L and N_U are the sample sizes smaller and larger than *CL*, respectively.

The lower bound *LB* and upper bound *UB* of the acceptable domain can be calculated by the following formulas:

$$LB = \begin{cases} CL - Skew_L \times \sqrt{-2\frac{\hat{S}_x^2}{N_L} \times \ln(10^{-20})}, & LB \le min \\ min, & LB > min \end{cases}$$

$$UB = \begin{cases} CL + Skew_U \times \sqrt{-2\frac{\hat{S}_x^2}{N_U} \times \ln(10^{-20})}, & UB \le max \\ max, & UB > max \end{cases}$$

$$Skew_L = \frac{N_L}{N_L + N_U}$$

$$Skew_U = \frac{N_U}{N_L + N_U}$$

$$\hat{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$$

Where $Skew_L$ and $Skew_U$ represent the left and right skewness of sample diffusion, respectively; $\hat{S}_x^2$ is the sample variance; *n* is the total number of samples.

Randomly draw n_v virtual samples according to a uniform distribution.

Calculate the membership function value *MF* of the observation point x_i to express the importance of the sample and the possibility of occurrence.

$$MF = \begin{cases} \frac{x_i - LB}{CL - LB} & x_i \le CL \\ \frac{UB - x_i}{UB - CL} & x_i > CL \end{cases}$$

MTD is a VSG technique developed and proposed based on information diffusion. The value of the triangular membership function is used to represent the likelihood of occurrence of sample points. The advantage of MTD to generate virtual samples lies in considering the integrity of the samples, taking the whole samples as the object of diffusion, making full use of the distribution trend of the samples, and deeply

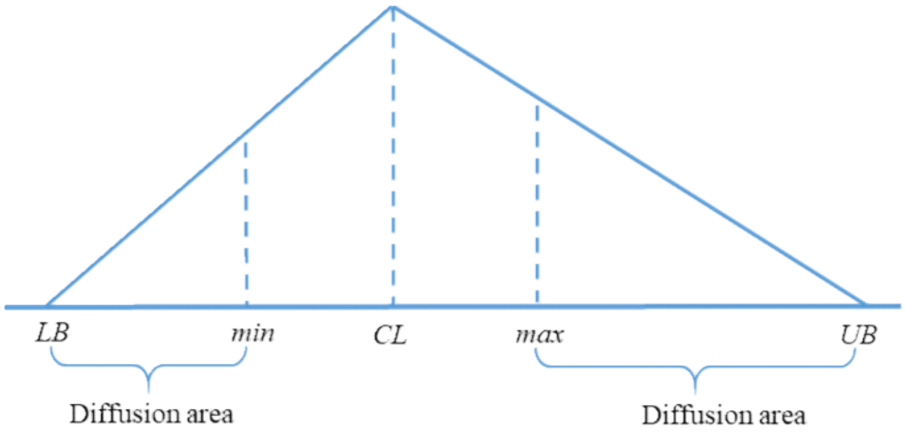

Figure 2. The scheme of MTD. MTD: mega trend diffusion.

exploring the useful information between the samples. However, the virtual samples generated by MTD are randomly drawn according to a uniform distribution, which may cause the obtained virtual samples to not conform to the distribution of the original samples. Due to the non-uniformity of realistic sampling, the values of N_L and N_U can be very different, resulting in unreasonable diffusion areas due to large left and right offsets, $Skew_L$ and $Skew_U$. Zhu *et al.* proposed a novel multi-distribution MTD technique by adding a correction quantity factor S_p on the calculation of $Skew_L$ and $Skew_U$ to prevent excessive $Skew_L$ and $Skew_U$ due to the non-uniformity of sampling[38]. The improved $Skew_L$ and $Skew_U$ formulas are shown below:

$$Skew_L = \frac{N_L}{N_L + N_U + S_p}$$

$$Skew_U = \frac{N_U}{N_L + N_U + S_p}$$

Yu *et al.* combined the advantages of both the MTD technique and the Monte Carlo method to develop the VSG technique of MC-MTD for generating effective, high-quality virtual samples to expand the original sample set to improve the learning ability and generalization ability of the model[39]. First, the MTD is used to estimate the acceptable range of each dimensional attribute of the sample, calculate the relevant data, and model the probability distribution of the sample according to the triangular affiliation function. Then, the Monte Carlo method is used for sampling to generate virtual samples. MC-MTD has been successfully applied to the multilayer ceramic capacitor (MLCC) dataset and the PTA dataset. The mean absolute percentage error of ten independent tests was reduced from 5.3% to 3.7% after adding 100 virtual samples using MC-MTD to the 25 initial samples of MLCC. The mean absolute percentage error of ten independent tests was reduced from 1.25% to 0.94% after adding 250 virtual samples using MC-MTD to the 25 initial samples of PTA.

Gaussian mixture model

As shown in Figure 3, a GMM is a probabilistic model that assumes that all data points are generated from a mixture of a finite number of Gaussian distributions[40,41]. If n observations $X = \{X_1, \ldots, X_n\}$ are generated by a mixture distribution P, where each vector X_i is p-dimensional, and the distribution P is composed of G components. Then the maximum mixture likelihood function of the distribution is shown in the formula:

$$L_M(\theta_1, \ldots \theta_G; \gamma_1, \ldots \gamma_n | x) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i|\theta_k) \left(\pi_k \geq 0; \sum_{k=1}^{G} \pi_k = 1\right)$$

Where $f_k(x_i|\theta_k)$ represents that X_i is the k-th density function; θ_k is the corresponding parameter; π_k is the weight parameter. If $f_k(x_i|\theta_k)$ is a multivariate normal distribution, then P is the Gaussian mixture distribution, where θ_k consists of the mean value μ_k and the covariance matrix $\sum_k$. The density function $f_k(x_i|\theta_k$ is shown in the formula:

$$f_k(x_i|\mu_k, \sum_k) = \frac{exp\left\{-\frac{1}{2}(x_i - \mu_k)^T \sum_k^{-1}(x_i - \mu_k)\right\}}{2\pi^{\mu/2}|\sum_k|^{1/2}}$$

The Gaussian mixture distribution P can be described by the probability density function represented by the weighted average of G Gaussian density functions, and the specific description is shown in the following formula:

$$P(x|\theta) = \sum_{k=1}^{G} \pi_k f_k(x_i|\mu_k, \sum_k)$$

A GMM is a density estimation algorithm that can be used to construct a probability model for small data. After calculating the parameters with the EM algorithm, the virtual samples that meet expectations could be generated through the model of the probability density distribution. Virtual samples generated by GMMs have been used for materials process optimization and industrial optimization. While GMM-based VSG techniques have shown success in material design and discovery, there is still potential for improvement in the GMM approach. One limitation is the potential presence of outliers in the virtual samples generated by GMM that may not accurately represent actual materials. Additionally, materials data often exhibit a range of eigenvalues, making it challenging to ensure that the generated data align with the characteristics of real materials. Currently, the analysis of anomalies in GMM-generated virtual samples relies heavily on empirical domain knowledge. However, in the future, objective analysis tools can be developed to better identify and analyze anomalies in these virtual samples.

In 2022, Shen *et al.* used a GMM combined with XGBoost to construct a machine learning model for predicting the wear resistance of rubber materials through mechanical properties[40]. The authors collected 24 rubber Acrolon abrasion test data as target property and corresponding six mechanical properties as descriptors from the publications and applied the algorithms of XGBoost, LASSO, support vector regression (SVR), and RF to construct the model. The results show that after generating 295 virtual samples with a GMM, the model constructed by XGBoost has the best prediction accuracy, with the R^2 of the test set reaching 0.95, which has been improved by 41% compared with the prediction accuracy before the original 24 samples. Our team also successfully designed the yttria-stabilized zirconia (YSZ) thermal barrier coating materials with high bonding strength using GMMs combined with machine learning[42]. First, eight experimental bonding strengths of YSZ thermal barrier coatings under four different atmospheric plasma spraying (APS) parameters were collected as the target property, and the corresponding APS parameters were taken as descriptors. Then, after expanding the data size from 8 to 400 with the GMM-VSG, the models were constructed and compared with various algorithms, including the ordinary least square (OLS), linear regression (LR), RF regression (RFR), decision tree regression (DTR), partial least squares (PLS),

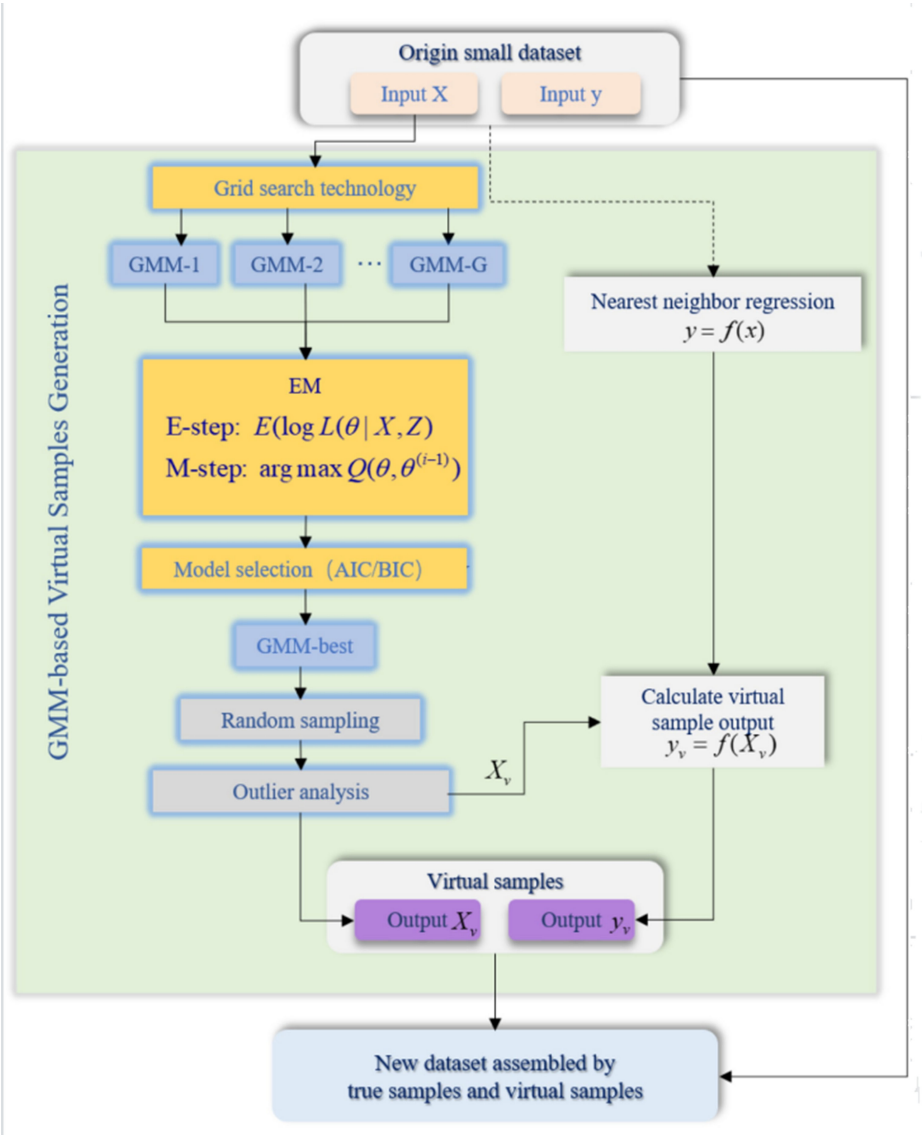

Figure 3. Flowchart of the proposed GMM-VSG. Reproduced with permission from ref.[40] Copyright 2022, Elsevier. GMM: Gaussian mixture model; VSG: virtual sample generation.

multiple LR (MLR), artificial neural network (ANN), and SVR with different kernel functions. The R and RMSE of leaving-one-out cross-validation (LOOCV) before and after data expansion by GMM are shown in Figure 4A and B. The results indicate that the SVR with polynomial kernel has the highest accuracy after a GMM generates virtual samples, and the R of LOOCV is as high as 0.989, while the R of LOOCV of the optimal ANN constructed by the original eight datasets could only reach 0.758. After sensitivity analysis and virtual sample analysis, we broke through the limit of the maximum bonding strength of 46.6 MPa in the original eight data and designed a sample with a predicted bonding strength of 53.137 MPa. After experimental validation, the experimental value of bonding strength has reached up to 55.5 MPa, and the absolute error with the model predicted value is only 2.363 MPa.

Random forest

RF primarily learns the distribution and features of samples by constructing multiple decision trees to

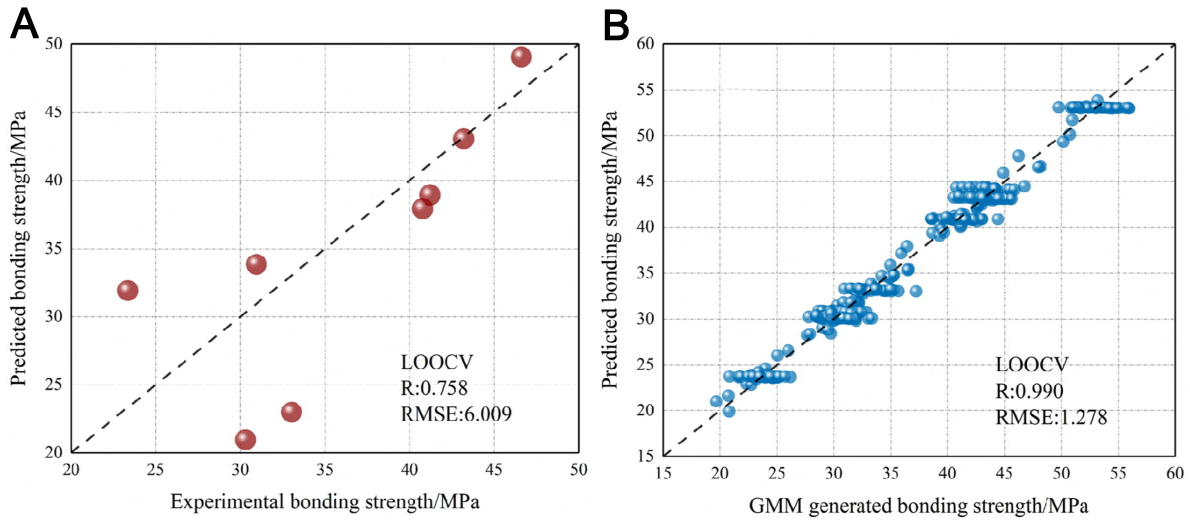

Figure 4. Experimental bonding strength *vs.* predicted bonding strength of LOOCV with corresponding R and RMSE based on (A) original data and (B) after data expansion by a GMM. Reproduced with permission from ref[42]. Copyright 2022, American Chemical Society. GMM: Gaussian mixture model. LOOCV: leaving-one-out cross-validation.

generate a new set of virtual samples[43,44]. In RF, multiple decision trees are combined for prediction and classification, with each decision tree constructed based on different random samples and features. Virtual samples are generated through the Bagging process in RF, where different random samples and feature sets are used to train models in each decision tree, resulting in the generation of diverse virtual samples[45]. The basic process of RF includes: (1) Using the RF model to train the original data; (2) Generating a specified number of virtual samples, where the target variable of each virtual sample is predicted by the RF model, and the feature is obtained by random extraction from the original data. The RF model is used to predict the target variable of virtual samples as it can effectively deal with high-dimensional features and nonlinear relationships. Generating features by random sampling of the original data can increase the diversity of virtual samples and improve the generalization ability of the model. However, the RF-based VSG may lead to a concentration of generated samples and a lack of diversity. More randomness can be introduced when generating virtual samples, such as introducing noise, random perturbation, or other data enhancement techniques to increase the diversity of the generated samples.

Generative adversarial networks

The VSG technology based on GANs is a deep learning technique that generates virtual samples similar to the original data through adversarial training[46]. A GAN comprises two neural networks: a generator and a discriminator. The generator generates virtual samples from random noise vectors, while the discriminator distinguishes between real and virtual data by detecting differences between them[46,47]. During the training process, the generator network continuously generates more realistic virtual samples, and the discriminator network continuously identifies differences between real and virtual data. This competitive process gradually leads to the generation of more realistic virtual samples by the generator network and improved accuracy of the discriminator network. Once both networks are optimally trained, the generator network can produce high-quality and diverse virtual samples, while the discriminator network can accurately differentiate real and virtual data. GAN technology has been widely applied in various fields, such as image, video, and audio generation[48,49]. In cases of insufficient data, GAN-based VSG technology can expand the dataset and enhance the performance and generalization ability of the model. However, the training process of GANs may not be stable enough, and the balance between the generator and discriminator is difficult to achieve, resulting in the generator generating low-quality samples or the discriminator not being able to

effectively distinguish between real and virtual samples. In order to improve the stability of training and the quality of generated virtual samples, different generator and discriminator architectures can be applied, such as deep convolutional generative adversarial networks (DCGAN), Wasserstein generative adversarial networks (WGAN), and conditional generative adversarial networks (CGAN), *etc*.

Zhu *et al*. proposed a CGAN-based VSG method (CGAN-VSG) to identify sparse regions of the data and generate the output of new virtual samples[50]. First, the local anomaly factor algorithm is used to identify discrete points in the dataset, which are also the potentially sparse regions. Then, the k-means++ is used to collect the discrete points to obtain the clustering centers[51]. Intermediate interpolation is performed between the centroids to generate target values for the virtual samples. The eigenvalues of the virtual samples are then generated using a CGAN. The datasets of two-dimensional criterion functions and three-dimensional (3D) criterion functions are used to verify the validity of the CGAN-VSG method. In addition, CGAN-VSG was used to predict the melt index (MI) for practical industrial applications in the production of high-density polyethylene (HDPE). After generating 120 virtual samples using five methods, including Bootstrap, MTD, *etc*., the MI prediction model for HDPE was developed by combining back-propagation networks. The comparison shows that the RMSE of the model trained by CGAN-VSG is the lowest at 0.4995, while the RMSE of the model without the VSG technique is as high as 0.5630.

SEARCHING ALGORITHMS FOR MATERIALS DISCOVERY

Searching algorithms for materials discovery essentially construct a virtual materials space combined with machine learning models to search for target materials. The process of constructing the virtual materials is the process of VSG. After encoding the materials into the combinations of segments or elements, a lot of virtual samples could be generated artificially. After obtaining the descriptors of the virtual samples, the target materials could be searched out by machine learning models. Commonly used searching algorithms for material discovery are presented, including particle swarm algorithms, efficient global optimization (EGO), and proactive searching progress.

Particle swarm optimization

The PSO algorithm is a population-based optimization algorithm to be used not only for generating virtual samples but also for materials searching[52-54]. The basic idea of PSO is to transform the optimization problem into a searching problem in a multidimensional space, where each materials sample is treated as a particle and the properties or structures to be searched are regarded as the objective function. The algorithm updates the position and velocity of each particle continuously to search for the optimal solution. The choice of the objective function should take into account the practical application requirements of the materials.

In the PSO algorithm, the particle swarm is composed of multiple particles, where each particle represents a material sample. Each particle has an initial randomly generated position vector and velocity vector, where the position represents the structural parameters of the materials and the velocity represents the change of the position. Based on the current position and velocity, the particle position is updated with the value of the objective function calculated using machine learning models. The fitness of each particle is then obtained based on the objective function, where higher fitness represents a performance indicator closer to the desired materials properties. The PSO algorithm is then used to update the position and velocity of each particle, taking into account the historical best solution of each particle and the entire particle swarm. The algorithm continuously calculates fitness to update the position and velocity, ultimately outputting the properties of the materials corresponding to the optimal solution in the particle swarm, which is the desired virtual material.

In materials searching, the PSO algorithm can be used to search for the optimal materials structure and virtual materials with target properties by continuously adjusting the particle position and velocity. Additionally, a multi-objective PSO algorithm can be used to transform multiple objective function optimization problems into a multidimensional space, thus searching for the structure of the optimal materials more comprehensively. With the PSO algorithm, we can quickly and efficiently search for the target structure and property of the virtual samples, providing more reliable theoretical support for materials design. Zheng *et al.* used the PSO algorithm to optimize the reflectivity band of absorptive materials and compared the bandwidth under different reflectivity conditions between the PSO algorithm and the genetic algorithm (GA)[55]. The study mainly aimed to optimize the bandwidth of materials with reflectivity less than -15 dB in the range of 2-18 GHz. It was found that the PSO algorithm had good frequency bandwidth under the conditions of reflectivity less than -10 dB and -20 dB, and corresponding dielectric constants, isolation layer thicknesses, and impedance values could be obtained. After three consecutive optimization searches, the bandwidths of two-layer absorptive materials with reflectivity less than -15 dB were compared. It was found that the PSO algorithm had a bandwidth of about 2 GHz larger than the GA and had better stability. Under the condition of reflectivity less than -20 dB, the bandwidths of two optimization algorithms were compared. The results showed that the bandwidth of the PSO algorithm was 3 GHz larger than that of the GA, and wider bandwidths could be obtained. Finally, under the constraint condition that the reflectivity of three-layer absorptive materials was less than -15 dB, the bandwidths of two optimization algorithms were compared to reveal that the PSO algorithm had a wider bandwidth and the sum of reflectivity was 2 dB less than that of GA. Overall, using the PSO algorithm to optimize the reflectivity band of absorptive materials has a good optimization effect.

Efficient global optimization

The EGO algorithm is a Bayesian optimization method that uses a Kriging surrogate model to predict the unknown objective function values and select the best parameter combination to optimize the objective function[56,57]. The Kriging surrogate model can be used to map the relationship between material properties and parameters[58]. By fitting the given data to the Kriging surrogate model, the EGO algorithm can predict the material properties of unknown virtual samples. In each iteration, the EGO algorithm would select the optimal parameter combination and generates new sample points near the combination to update the surrogate model. This strategy can help the algorithm converge quickly to the global optimal solution. The EGO algorithm has wide application potential in materials design and material searching fields[59]. By modeling and optimizing the objective function, the EGO algorithm can quickly search for materials with excellent performance with fewer computational resources.

After the target properties and features of material design are determined, the EGO algorithm initializes a search space for materials composition and structure based on these features. Next, sampling points would be selected in the search space of virtual samples with their corresponding objective function values calculated by the surrogate model. Then, based on the objective function values of the sampling points, a surrogate model is fit, and the next sampling point is selected. Common selection strategies include confidence intervals and the expected improvement (EI) strategy. Through iterations of selecting sample points and fitting the Kriging surrogate model, the algorithm would keep operating until it reaches a preset stopping criterion, such as the maximum number of sampling points or convergence of the objective function value. Finally, by analyzing the relationship between the objective function values of the sampling points and the material parameters, the EGO algorithm finds the optimal material parameter combination and carries out material preparation and property testing validation.

The EGO algorithm uses the Kriging surrogate model to fit and predict the objective function, enabling it to quickly search for materials with excellent performance with fewer computational resources, and has the advantages of high efficiency and high accuracy. It can also improve search efficiency by introducing multiple surrogate models and has good scalability. However, if the sample dimension in the search space is too high, the search efficiency of the EGO algorithm may be limited, and the optimization results of the EGO algorithm may be affected by the initial sampling points.

In addition to being used for target materials searching, the EGO algorithm can also be combined with an active learning framework. In the active learning process, the core step is to sample important samples from the unlabeled sample pool, and the EGO algorithm can accomplish this task. Zhao *et al.* combined the EGO algorithm as a representative sampling strategy with an active learning strategy to develop an effective machine learning model for predicting the hardness of 6061-aluminum alloy elements[60]. First, they used a full-process high-throughput alloy preparation and characterization system to prepare 32 different composition ratios of 6061-aluminum alloys and characterized their hardness. The initial 309 descriptors were constructed by the composition of elements and the knowledge of the alloy field. After feature selection, the remaining five important features were used for modeling. After comparing multiple algorithms, the SVR algorithm with a radial basis kernel function was used to construct the aluminum alloy hardness prediction model. Artificial experience sampling and Bayesian optimization sampling were used to select samples from candidate materials for labeling and subsequent experiments. The Bayesian sampling strategy used four methods, the EGO algorithm, the knowledge gradient (KG) algorithm, the maximum hardness point method, and the maximum error point method, with four data points for each method and a total of 16 experimental alloy compositions were designed for the next round of experiments. Before reaching the convergence condition, the experimental data was returned to the initial dataset for further feature selection and model construction. After three iterations, the results indicated that the adaptive sampling strategy based on Bayesian optimization can more effectively guide experiments than artificial experience sampling, with a 63.03% decrease in an average absolute error (MAE) and a 53.85% decrease in RMSE. The RMSE was 4.49 HV, which was close to the experimental error of 4.05 HV for the test samples. Xue *et al.* collected 22 Ni-Ti-based shape memory alloys and used atomic parameters as descriptors and thermal hysteresis as the target variable to establish a hysteresis prediction model using the SVR algorithm[61]. The EGO algorithm was used to search for four samples with low hysteresis from a search space of 800,000 samples, which were then synthesized and characterized through experiments. After experimental validation, the four samples were put back into the training set for iterative modeling, searching, and experimentation. After nine iterations, a total of 36 samples were searched, of which 14 had lower hysteresis than any of the 22 samples in the original dataset. The $Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$ sample had the lowest hysteresis, which was only 1.84K.

Proactive searching progress

The proactive searching progress (PSP) is a materials search method developed based on the Sequential Model-Based Optimization (SMBO) method[62,63]. Its core idea is to treat the materials composition to be optimized as parameters and use a lower-cost proxy model to learn the relationship between the existing parameters and properties, and quickly search for the approximate trend of the optimized composition in the chemical space. The PSP method uses a pre-constructed high-precision machine learning model to predict the precise performance of the optimized composition, and after repeated iterations, it can gradually approach the optimized composition o_* with the desired performance[64].

Assuming that a materials composition space is constructed as $\{\boldsymbol{\gamma}_i \mid \boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots\gamma_\alpha)^T, i, d \in R\}$, where i and d are positive integers, T represents transpose, and $\boldsymbol{\gamma}_i$ is any materials composition represented by a vector. $\gamma_{i1}, \gamma_{i2}, \ldots \gamma_{id}$ are the specific components of the materials composition $\boldsymbol{\gamma}_i$. Based on this materials composition

space $\{\gamma_i \mid i \in R\}$, the pre-constructed high-precision machine learning model can be used to predict the performance values $\{o_i \mid i \in R\}$ of each composition in this space, forming a set of compositions and performance $\{(\gamma_i, o_i) \mid i \in R\}$. In this process, the pre-constructed high-precision machine learning model actually plays the role of a function mapping between the composition and the performance, that is, $o = f(\gamma)$, which can be used to describe the distribution of materials performance samples in the global materials composition space.

Combining the idea of SMBO in machine learning model parameter optimization, a proxy model can be constructed to search for optimized compositions with desired properties. In order to shift the materials composition exploration approach from traversal to search, the function mapping relationship between the composition and the performance $o' = g(\gamma)$ constructed by the proxy model requires continuity and differentiability so that its derivative $g'(\gamma)$ can be used to determine the direction of the predicted performance o', in order to meet the necessary conditions for search. Since the GPR algorithm has the characteristics of a simple algorithm, parameter-free modeling, and high accuracy, it can be used to construct the proxy model.

The process flowchart for the PSP method is shown in Figure 5. In the first step, several materials components $\{\gamma_i \mid i \geq 2\}$ are randomly generated from the existing performance distribution in the materials composition space. The number of materials components should be greater than or equal to 2, and their performance is predicted using a high-precision machine learning model to form the dataset $\{(\gamma_i \mathrm{EE}_i) \mid i \in R\}$ for fitting the surrogate model. In the second step, a GPR model is built based on the existing dataset $\{(\gamma_i, EE_i) \mid i \in R\}$. In the third step, the materials composition that minimizes EE_{GP} is found in the GPR model and predicted using a high-precision machine learning model. If the prediction error is less than a pre-defined threshold, the sample is outputted. Otherwise, it is added to the existing dataset $\{(\gamma_i, EE_i) \mid i \in R\}$, and a new GPR model is constructed for the next round of search.

Lu *et al.* of our team used integrated machine learning techniques and the PSP method to predict the band gaps of hybrid organic-inorganic perovskites (HOIPs)[64]. The author collected 1,201 samples from publications with 129 atomic descriptors, including atomic radius, electronegativity, tolerance factor, tao factor, and octahedral factor. Then, various ML algorithms were used to construct models. The top 4 models (including CatBoost, XGBoost, LightGBM, and Gradient Boosting Machine) were selected and integrated using weighted voting regression to achieve better performance. The weighted voting regressor (WVR) model achieved R^2 and RMSE of 0.95 and 0.079 in LOOCV and 0.91 and 0.106 in the test set. Based on the ions collected from the formulas in the dataset, the author constructed a vast material space, including over $8.20*10^{18}$ possible combinations, to explore new HOIP structures with suitable band gaps. As shown in Figure 5, the PSP method effectively searched for material combinations with expected band gap values from the generalized chemical space. As a result of the PSP method, 20,242, 733,848, 764,883, and 746,190 lead-free candidates were designed for HOIPs with band gaps of 1.20, 1.34, 1.70, and 1.75 eV, respectively. To validate the PSP searching results and the prediction ability of the WVR model, the author synthesized new HOIP compositions $MASn_xGe_{1-x}I_3$ (x = 0.85, 0.74, 0.66) and conducted experimental validation, with an average error between the experiment and prediction of only 0.07 eV. Although the PSP method enables rapid exploration of a vast space of material components to identify target materials with desirable properties, the SMBO search method employed in PSP can be substituted with alternative search methods. By utilizing different search techniques, not only can diverse search requirements be fulfilled, but also the search efficiency can be further enhanced. Some commonly employed search methods include gradient descent and Newton-like methods.

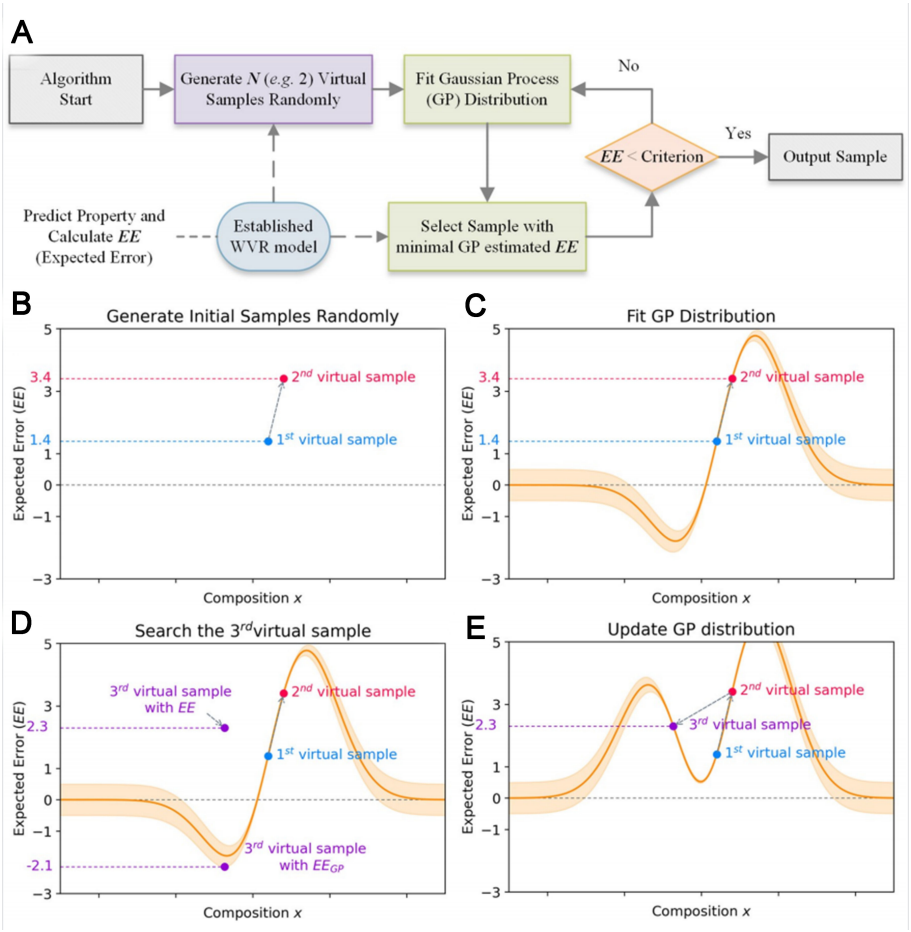

Figure 5. General principle of the PSP method applied for searching HOIPs with targeted band gaps. Reproduced with permission from ref[64]. Copyright 2022, American Chemical Society. HOIPs: Hybrid organic-inorganic perovskites; PSP: proactive searching progress; WVR: weighted voting regressor.

MATERIALS INVERSE DESIGN ALGORITHMS

Machine learning models for the prediction of material properties mainly focus on the mapping from the features to the properties of the materials. In contrast, the inverse design of materials requires a mapping from properties to features, which aims to generate the virtual samples by the requirement of the target properties and the machine learning model. Different from searching algorithms for materials discovery, inverse design does not require the design of a large number of virtual samples for searching. The inverse design is oriented to the target property of the materials to generate virtual samples that satisfy the target properties directly through the principle of algorithms. Commonly used algorithms for materials inverse design include GA, Bayesian optimization, and pattern recognition inverse projection.

Genetic algorithm

GA is a computational method that simulates the process of biological evolution, following the principle of natural selection described in Darwin's theory of evolution[65,66]. By simulating basic genetic operations, GA can seek optimal or near-optimal solutions to problems. Material inverse design based on GA uses computer simulation and optimization to design virtual materials, with a focus on the materials encoding and evolutionary mutation[67,68]. In GA, material structures are encoded as a string, such as a binary string or a string of characters. For each structural parameter of the materials, such as lattice constants or atomic

positions, a binary digit or character is used for encoding. Materials encoding is more focused on digitizing and compiling the composition and structural information of materials, which is different from material descriptors. For example, organic structures can be encoded using a simplified molecular input line entry system (SMILES)[69]. Encoding molecules in the SMILES format can effectively save storage space and has been widely used in chemical databases and data-driven molecular design. The evolution process of GA mainly includes selection, mutation, and crossover to generate virtual samples. Through the evaluation and screening of newly generated virtual samples using fitness functions, the virtual samples with high fitness could be left behind and put into the next generation for evolution. The materials inverse design strategy based on GA includes the following steps:

Design of initial population: The initial population is the starting point in the materials inverse design process. It can be generated by random generation or by mutation and recombination based on existing materials.

Definition of fitness function: The fitness function is the criterion for evaluating the quality of each individual. Depending on the characteristics of the materials, goals, and constraints of the inverse design, a fitness function can be constructed. The fitness function is usually the error between the expectation and prediction via the machine learning model.

Execution of genetic operations: Selection, crossover, and mutation are the three basic operations of GAs. These operations can produce new individuals, and selection is performed based on the fitness function in each iteration to evolve more suitable individuals.

Setting the end condition: The inverse design process can be iterated multiple times until the set end condition is reached, such as reaching the maximum number of iterations or the fitness reaching a certain threshold. Materials structures that meet the fitness function are decompiled and outputted.

The GA has been successful in the applications of materials inverse design. For example, the highly efficient mechanical performance of minimal surface structures has drawn widespread attention from scholars due to their lightweight and high strength. However, the existing research has mainly focused on the mechanical properties of different minimal surface configurations. Therefore, the inverse design of surface-like metamaterial configurations that meet the requirements based on mechanical properties is still a research gap. Wang *et al.* proposed a GA-based algorithm that combines efficient machine learning methods and globally optimal solutions to achieve the inverse design of mechanical metamaterials based on load curves[70]. To generate a dataset for constructing the machine learning model, a numerical model with different mechanical metamaterial configurations was constructed through the finite element method (FEM) to predict the load curve of the shell-based mechanical metamaterials (SMMs). A total of 7,000 SMM configurations were generated, with 5,500 used for the training set and 1,500 used for the test set. ANNs were used to map the relationship between the load curve and the geometric configuration to achieve forward prediction of the load curve through geometric configuration. The GA was used to search for the SMM configuration that most closely matched the target load curve. The flowchart of the GA-based algorithm for the inverse design of SMMs is shown in Figure 6. The individuals and the chromosomes in a GA are SMM configurations and geometric parameters, respectively. A total of 10,000 random samples were generated as the initial population. 10,000 and 1,000 samples were generated through crossover and mutation, respectively. Finally, 21,000 samples were selected, with 10,000 samples selected as the next population based on the load curve calculated by the previously trained ANN model. The results showed that regardless of considering strain hardening or softening, SMM configurations that closely match the

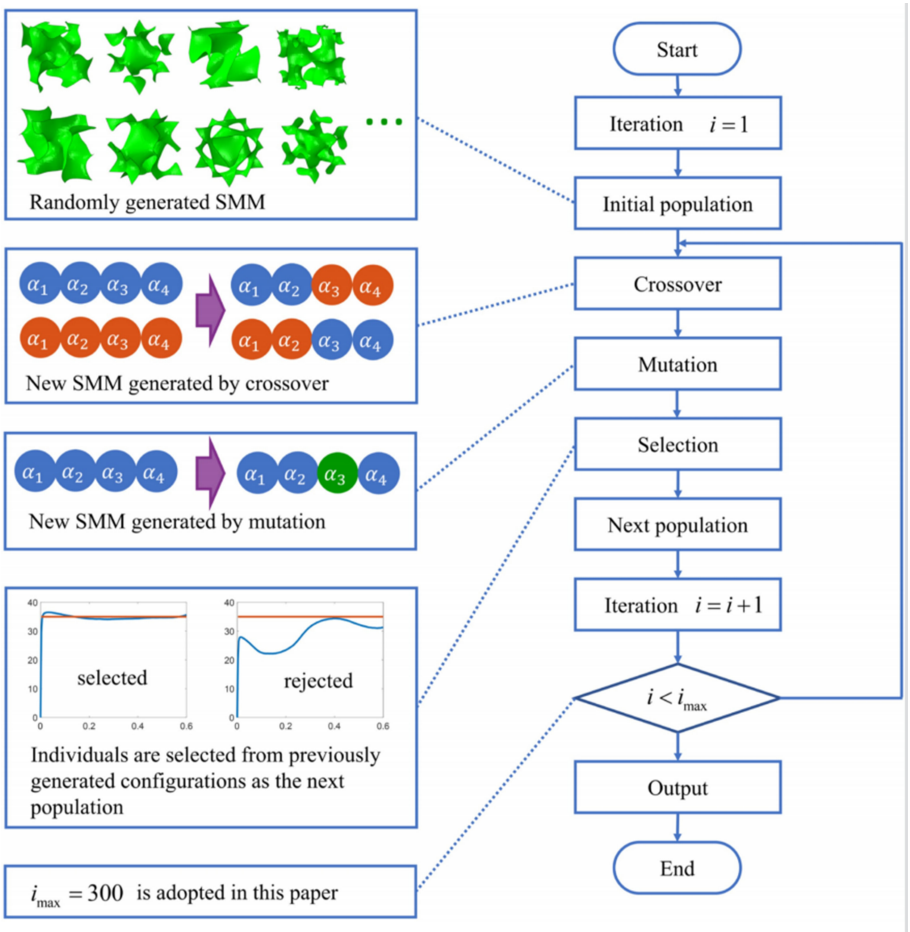

Figure 6. The flowchart of the genetic algorithm for the inverse design of SMM. Reproduced with permission from ref[70]. Copyright 2022, Elsevier. SMM: Shell-based mechanical metamaterial.

target load curve can always be obtained through GA inverse design. To explore the relationship between the load curve and the deformation mode, the authors designed several SMM configurations based on different types of load curves and then simulated the deformation mode under 4*4*4 unit compression using the FEM. The results showed that single cells with a "hardening" load curve tend to undergo overall deformation during the macroscopic overall structural deformation process, while those with a "softening" load curve tend to undergo layer-by-layer deformation during the macroscopic overall deformation mode. This work has filled the gap in the inverse design of SMMs and can be used to design SMM materials with specific load curves. It also contributes to the structural design concept that combines machine learning and traditional optimization methods.

Maurizi *et al.* combined machine learning methods with GAs to construct a composite model for the inverse design of non-uniformly assembled lattice materials, with flexural strength as the target variable[71]. Deep neural networks (DNNs) were used to construct the model to predict the yield strength to optimize the desired lattice architecture. Then, the GA was used to combine simple and diverse basic building blocks into various architectural itineraries to design spaces with DNNs to explore the design space to find the optimal architecture. For the GA-based inverse design, the fitness function is set to consider the average flexural strength and flexural isotropy in different load directions. Given the random nature of the GA algorithm, 100 iterations of optimization rounds were conducted until 12 different candidate designs were

obtained. The best performing six design candidates were selected after ranking the 12 candidates based on their fitness values. Three-dimensional printing was used to obtain samples of the designed lattice architectures for mechanical testing, and the test results were consistent with the simulation results, with flexural properties 30%-90% and 10%-30% higher than those of conventional reinforced lattice-like lattices and bionic lattices, respectively. The composite model proposed in this work for the inverse design of material structures can autonomously select basic components to construct more complex primitives to arrange into periodic structures, fully exploiting the periodic and local properties. Based on this composite model, truss grids with better buckling resistance than conventional reinforced lattices or bionic sponge constructions are obtained.

Nigam *et al*. proposed JANUS, a GA based on the SELFIES representation of molecules[72]. JANUS consists of iterations triggered by random structures or provided molecules. In each iteration, two fixed-size molecular populations are maintained. Members of each population compete with each other to enter the next generation. In a mutation and crossover design, JANUS relies on the STONED algorithm for the efficient generation of molecules. In terms of selection pressure, JANUS trains a DNN model for each generation to accurately predict the fitness function and uses a classifier to classify high-fitness samples from low-fitness samples. The trained DNN model evaluates the generated offspring samples and adds the highest fitness samples to the population. Inspired by parallel tempering, JANUS maintains two populations that use different genetic manipulations. One population performs a local search of the chemical space, using molecular similarity as a selection pressure. The other performs a global exploration using DNNs as selection pressure. JANUS has been successfully applied to maximize the performance of penalty logarithms for octanol-water partition coefficient scoring, protein inhibitor design, and docking scores of protein targets in molecular docking. The successful application of JANUS in the above molecular design cases also demonstrates the great potential in materials inverse design.

Bayesian optimization

Bayesian optimization is an optimization method based on the Bayes theorem. By combining prior knowledge and new observations and continuously updating the posterior distribution, it finds the global optimal or approximate optimal solution[73]. Compared to traditional optimization algorithms such as gradient descent, the Bayesian optimization algorithm has better performance in finding global optimal or approximate optimal solutions, especially in high-dimensional, noisy, or non-differentiable situations[74]. Therefore, it has been widely used in many application fields, such as materials inverse design and hyperparameter optimization.

In the field of materials inverse design, Bayesian optimization follows the following Bayesian theorem[75]:

$$p(S|Y \in U) \propto p(Y \in U|S)p(S)$$

Y, U, and S represent material properties, target material properties, and material structure encoding, respectively. S can also be encoded as a SMILE file. $p(Y \in U \mid S)$ represents the probability of target properties given a known structure, which can be obtained by training a machine learning model with structure-property relationship data. The prior distribution $p(S)$ is the probability of constructing a structure, which is used to reduce the occurrence of invalid or unstable chemical structures by setting zero or lower probability mass for them. According to the above Bayesian law, as long as $p(Y \in U \mid S)$ and $p(S)$ are obtained, the probability of structure appearance under the target property $p(S \mid Y \in U)$ can be obtained, thereby achieving the inverse design of the materials.

The core of Bayesian optimization can be divided into a generator and an evaluator. The generator refers to the posterior distribution $p(S)$ generated by the algorithm, where $p(S)$ is the most critical factor that affects the structural features of the generated samples, and its calculation formula is as follows:

$$p(S) = p(s_1)\prod_{i=2}^{g} p(s_i|s_{i-1}, \ldots, s_1)$$

Where s_i represents the i_{th} block that forms the materials. The probability of the i_{th} letter appearing depends on the previous $s_{i-1},\ldots s_1$. As mentioned above, the probability of unstable chemical structures can be effectively reduced by observing the records of subsequent characters given a given substring. The generator could be constructed with the extended n-gram model consisting of a table to record the probability of observing a subsequent character given a substring and a function to modify a given SMILES string based on the stored n-gram probability table. The evaluator refers to the likelihood function that evaluates the fitting degree of S with the attributes, which is the process of obtaining $p(Y \in U \mid S)$ by constructing machine learning models, defined as follows:

$$p(Y \in U \mid S) = \int_{U} \prod_{i=1}^{m} \frac{1}{\sigma_i(\varphi(S))\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \mu_i(\varphi(S))}{\sigma_i(\varphi(S))}\right)^2\right) dY_1 \ldots dY_m$$

The term $\varphi(S)$ refers to the material descriptor. The property of the i_{th} sample Y_i is a function of the descriptor $\varphi(S)$. The σ_i $(\varphi(S))$ and $\mu_i(\varphi(S))$ are the mean value and standard deviation of property Y_i, respectively, predicted by the forward machine learning model. With known $p(S)$ and $p(Y \in U \mid S)$, the posterior model $p(S \mid Y \in U)$ can be obtained, thus enabling the design of material structures with ideal performance.

As shown in Figure 7, Wu *et al.* developed a Bayesian-based inverse molecular design algorithm called iQSPR-X, which has been integrated into the materials informatics platform XenonPy, and successfully used to design polymers with specific band gaps and dielectric constants[75]. First, the authors collected 854 polymer structures composed of nine types of atoms from the Polymer Genome (PG) database and used the hybrid Heyd-Scuseria-Ernzerhof (HSE06) exchange-correlation functional to calculate their band gaps and the density functional perturbation theory (DFPT) to calculate the sum of electronic and ionic dielectric constants. Two strategies were considered for training the generator: one was to use the 854 samples collected from PG as the training set, and the other was to add samples collected from PubChem to the PG samples as the training set. Since polymers containing F atoms generally have high band gaps and dielectric constants, 2,485 samples were collected from PubChem, with at least six F atoms in one molecule. The results showed that the generator trained on PubChem data would contain more structures with F fragments during the molecule modification process. In the training of the estimator, a well-trained neural network was first used to evaluate ten descriptors. The results showed that the atomic pair fingerprint with MACCS keys exhibited high performance in predicting band gaps and dielectric constants and was, therefore, selected for Bayesian molecular design. Gradient boosting, RF, and Bayesian LR were used to establish prediction models for band gaps and dielectric constants by combining the optimal descriptors selected by the estimator. The results showed that the gradient boosting model had the best overall performance, with MAEs of 0.320 and 0.142 for band gaps and dielectric constants, respectively, in 5-fold cross-validation, and was, therefore, selected as the optimal algorithm for inverse design. The goal of this

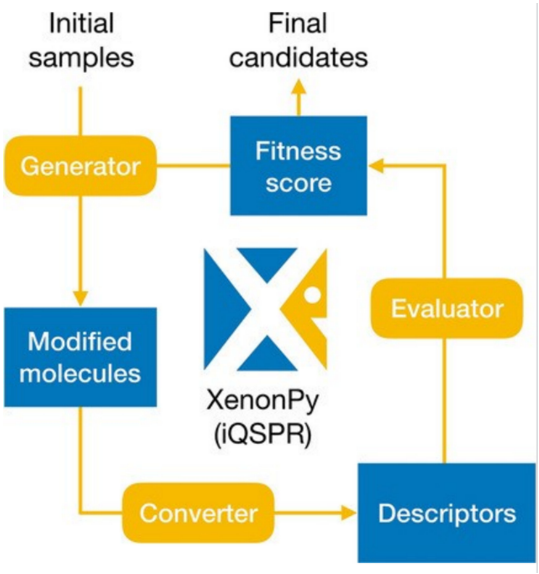

Figure 7. Computational workflow in iQSPR-X. Reproduced with permission from ref[75]. Copyright 2020, John Wiley and Sons.

work was to design polymers that simultaneously satisfy high band gaps and high dielectric constants, and three different strategies were used for inverse design. Strategy (1) trained the generator using the samples in PG and randomly selected 100 samples as initial samples for inverse design. Strategy (2) trained the generator using the samples in PG and selected 100 initial samples with low band gaps and dielectric constants ($\varepsilon < 4$ or $E_g < 4.5$ eV) for inverse design. Strategy (3) trained the generator using the samples from PG and PubChem and randomly selected 100 samples from the 854 samples as initial samples for inverse design. The results showed that different strategies led to the design of different polymers. Strategy (2) had difficulty converging to candidate molecules similar to those of strategy (1) and showed difficulties in capturing trends in complex ring structures. Strategy (3) showed a clear trend of attaching F fragments to the chemical structure.

This team also used transfer learning and Bayesian molecular design algorithms to develop a highly accurate prediction model for polymer thermal conductivity and screened thousands of polymer materials with high thermal conductivity[76]. After screening and experimental validation, three candidate samples successfully passed the experimental evaluation. The authors collected a large number of polymer samples from the PolyInfo and QM9 databases, including melting point T_m, glass transition temperature T_g, density ρ, and heat capacity C_p and C_v data, for pre-training model construction. After model comparison, the pre-training model constructed with heat capacity C_v performed the highest prediction accuracy. Next, the parameters of the pre-training model were optimized using 28 samples with thermal conductivity data to be transferred to the prediction of thermal conductivity. The results showed that the transferred model had an MAE of 0.0204W $(m \cdot k)^{-1}$, which was 40% lower than the direct training model using 28 data points with RF. Combining with the Bayesian algorithm to design a large number of polymers repeating unit structures, the screening was based on whether the repeating unit contained a liquid crystal structure, experimental feasibility evaluation, glass transition temperature, and the predicted thermal conductivity from the transferred model. Finally, 24 molecular structures were selected, three of which were successfully validated through synthesis and characterization experiments. The experimental results showed that the thermal conductivity of the polymers screened with the assistance of transfer learning and Bayesian molecular design was higher than that of polymer materials published in previous papers. This research also confirms the successful application of transfer learning and Bayesian molecular design in the design and discovery of polymer materials.

Pattern recognition inverse projection

Pattern recognition is a technique used to analyze and process data or signals to identify or classify different patterns or features. With the applications of computer technology, pattern recognition can extend features to a mathematical model of the distribution range of various samples in multidimensional space. Common pattern recognition methods include principal component analysis (PCA) and linear discriminant analysis (LDA).

PCA is a widely used linear dimensionality reduction technique that maps high-dimensional datasets into low-dimensional space while preserving the maximum variance information of the original data[77]. The basic idea of PCA is to project the original dataset onto a new coordinate system through a linear transformation so that the projected data has the maximum variance to extract the most representative principal components. LDA is a classic pattern recognition method used to divide datasets into different categories and perform classification[78]. The basic idea is to reduce the dimensionality of the data while maximizing the differences between different categories and minimizing the differences within the same category, thereby achieving effective data classification.

Pattern recognition techniques represented by PCA could compute two principal components linearly combined by descriptors, projecting the material samples onto a plane composed of different principal components for visualization to draw an optimization area according to the projection of samples and obtain the boundary equation of the optimization area. Sampling is performed in the optimization area, and the descriptor values could be calculated according to the PCA equation and deduced to obtain the chemical formula of the materials.

Pattern recognition inverse projection tends to be used in conjunction with machine learning to achieve performance breakthroughs in alloy materials. Yang *et al*. proposed a machine learning-based alloy design system that integrates database construction, model construction, composition optimization, and experimental validation to guide the rational design of high-entropy alloys with high hardness[79]. Using pattern recognition inverse projection technology, they successfully designed samples with hardness beyond the original dataset and verified them with experiments. First, they collected 370 samples of high-entropy alloy compositions and Vickers hardness information from the publications. After screening using the Pearson correlation coefficient, RF feature importance ranking, forward feature selection, and best subset method, a high-entropy alloy hardness prediction model was constructed using five important descriptors combined with SVR. The R values of the independent test and LOOCV both reached 0.94. Based on the model, the pattern recognition inverse projection method was adopted to explore the optimized composition of high-hardness high-entropy alloys, as shown in Figure 8A. The inverse projection method can obtain the features of the design samples in the original space, thus obtaining the corresponding compositions of the two designed samples, which are $Co_{18}Cr_7Fe_{35}Ni_5V_{35}$ and $Al_{20}Cr_5Cu_{15}Fe_{15}Ni_5Ti_{10}V_{30}$. Using the constructed SVR model to predict the hardness of the design samples, the predicted results were 1,002 HV and 1,028 HV, respectively. The predicted hardness of both optimized samples exceeded the hardness of the highest alloy in the original dataset. These optimized samples have the potential to be high-hardness high-entropy alloy compositions. They synthesized these high-entropy alloy samples using vacuum arc melting to measure the experimental hardness. The results showed that the $Co_{18}Cr_7Fe_{35}Ni_5V_{35}$ alloy has ultra-high Vickers hardness, reaching 1,148 HV, which is 24.8% higher than the highest hardness alloy in the original dataset.

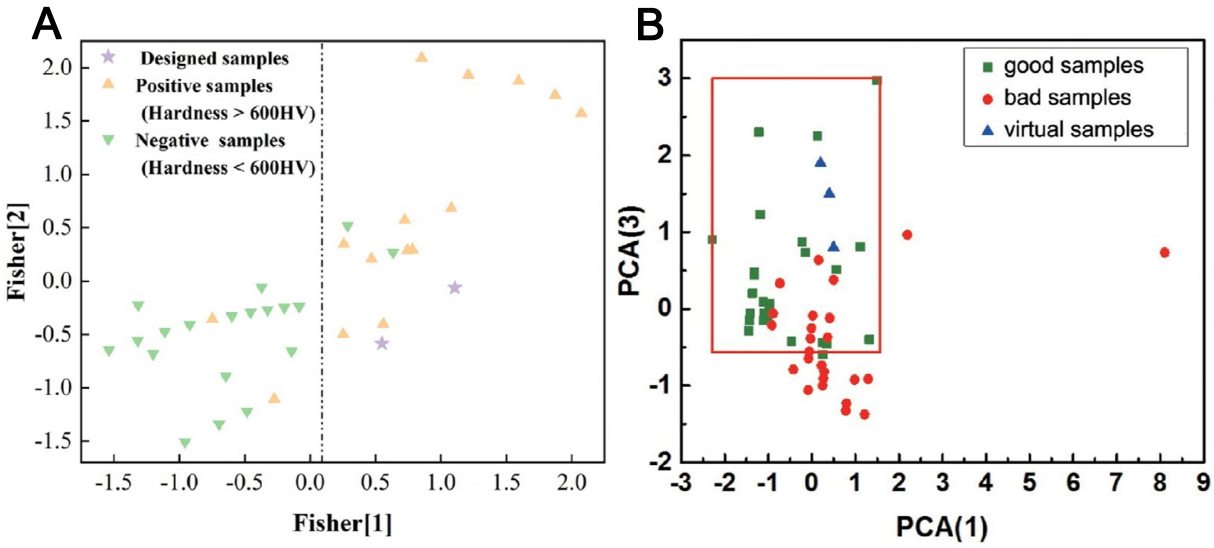

Figure 8. Materials pattern recognition of different samples by (A) Fisher and (B) PCA. Reproduced with permission from ref[79]. Copyright 2021, Elsevier. Reproduced with permission from ref[80].Copyright 2021, Chinese Materials Research Society. PCA: principal component analysis.

Gold and its alloys, especially ternary gold alloys, have been widely used in the field of electrical contact materials. Due to the complexity of the composition and proportion of ternary gold alloys, designing ternary gold alloy electrical contact materials with low electrical resistance is still a challenge. Wang *et al.* proposed an inverse design method for low electrical resistance ternary gold alloys, which combines pattern recognition inverse projection with XGBoost to design new materials with lower resistivity than existing ternary gold alloys[80]. First, the authors collected 51 ternary gold alloy samples at room temperature and pressure from the Materials Platform for Data Science (MPDS) database, including 62 atomic parameters and two-component features. After screening variables using Pearson correlation coefficients and maximum relevance minimum redundancy, five important variables were used for subsequent modeling and inverse design. Using the PCA method to establish an optimization area for low resistivity in Figure 8B, three points were selected as virtual samples in the best pattern recognition projection. Then, the pattern recognition inverse projection method was used to calculate the feature variables of the three virtual samples. Finally, by calculating the Euclidean distance, candidate samples closest to the virtual sample points were obtained, which were $AuZr_{1.95}Cu_{0.52}$, $AuZr_{1.12}Cu_4$, and $AuSc_{1.86}Cu_{2.75}$. XGBoost, SVR, multiple LR, and ridge regression were used to construct a prediction model for the electrical resistivity of gold alloys. The results showed that the XGBoost model had the highest R value and the lowest RMSE value, which were 0.850 and 0.331, respectively. The XGBoost predicted resistivity values for the three candidate materials were 6.718, 6.707, and 6.701, respectively, all of which were higher than the maximum value of 6.68 in the original dataset. Therefore, pattern recognition and its inverse projection algorithm can be used for the inverse design of low electrical resistance ternary gold alloy materials.

DISCUSSION

It can be found that although there are many VSG algorithms for data expansion, not every algorithm has been successfully applied in the materials field. Based on our previous work, we utilized eight datasets along with VSG techniques of RF, MTD, and Bootstrap to increase eight samples to 400. These 400 virtual samples were then divided into a 4:1 ratio for training and testing, with 320 samples in the training set and 80 samples in the test set. To ensure comparability, we used a support vector machine with a polynomial kernel function that was previously described in the reference as the modeling algorithm. The evaluation functions were R and RMSE of LOOCV, 10-fold CV, and the test set. The modeling results, as presented in Table 1,

Table 1. R and RMSE of LOOCV, 10-fold CV, and the test set with different VSG methods

	RF	MTD	Bootstrap	GMM
R_{LOOCV}	-0.235	0.493	0.989	0.990
$RMSE_{LOOCV}$	4.884	6.596	1.164	1.278
$R_{10\text{-fold CV}}$	-0.154	0.482	0.989	0.990
$RMSE_{10\text{-fold CV}}$	4.871	6.644	1.164	1.280
R_{test}	0.114	0.613	0.987	0.986
$RMSE_{test}$	4.553	5.710	1.164	1.416

CV: Cross validation; GMM: gaussian mixture model; LOOCV: leaving-one-out cross-validation; MTD: mega trend diffusion; RF: random forest; RMSE: root mean square error; VSG: virtual sample generation.

indicate that Bootstrap yielded similar results to GMM with much better evaluation parameters than RF and MTD. Consequently, our modeling results have confirmed that Bootstrap is the superior method. Nevertheless, Bootstrap relies on repeating sampling, resulting in numerous duplicate samples in the generated virtual samples. Additionally, Bootstrap did not explore the distribution information of the data or make up for information deficiencies. Although our modeling result was optimal, its practical value was less than GMM. There is currently a debate over the use of VSG to increase the number of training set samples, as it contradicts empirical knowledge that VSG can significantly enhance prediction accuracy. While the published papers have proven that VSG technology has been partially successful in materials machine learning, some VSG methods, such as GMM, still lack sufficient theoretical support in practical applications. Assigning a Gaussian distribution to data and generating virtual samples on the Gaussian distribution is rather controversial. This contradiction between machine learning results and domain knowledge is one of the three contradictions in materials machine learning[81]. Presently, although VSG technology has shown some success in materials design and discovery, more rigorous theoretical support is still necessary, such as evaluating different VSG methods using the same data, exploring the number of virtual samples generated, and investigating the influence of the VSG method parameters.

The most challenging aspect of using searching algorithms to discover materials with target properties is the construction of the materials space. In the PSP searching, materials components are set as a combination of elements and doping ratios. The chemical formula of the sample is unified into "element-doping ratio-element-doping ratio-..." to eliminate the interference of samples with identical elements and doping ratios but only in different element order, which limits the scale of descriptors to the component descriptors. For organic materials, the applications of the searching algorithms for materials discovery are very limited because the influence of the 3D structure on the properties should be taken into account; structural optimization is often required, and the descriptors may contain much more information. In addition, the simple setting of material components as a combination of elements and doping ratios is very beneficial to the construction of materials space, but the huge materials space constituted by the diversity of element types and doping ratios also contains a lot of redundant information, such as structures that would not exist or be unstable. Therefore, it is necessary to filter out the redundant information and search for the target materials by using domain knowledge to restrict the materials space before construction or by using other criteria after construction. For example, in the case of doped ABO_3 perovskites, there are various choices of doping elements and ratios for A- or B-sites, but not all dopants can form perovskite structures. Therefore, after constructing the materials space, the tolerance factor or octahedral factor can be used to eliminate the combinations that cannot form perovskites before searching.

For the inverse design of materials, the most challenging aspect is the correspondence of the materials encoding-descriptor-property relationship. If the improvement of material properties is only through the

optimization of experimental parameters, the research task could be simplified with only experiments conducted according to the optimized parameters. However, if the optimization of material components is used to improve material properties, the relationship between materials encoding, properties, and descriptors must be matched one by one. Different encoding rules for the same materials combined with descriptor generators and algorithms can lead to models with different accuracy to achieve forward prediction from materials encoding to descriptors and properties. The difficulty of inverse design is to invert the descriptors from the properties to the materials encoding. Moreover, in the process of machine learning, feature selection is often performed to improve model accuracy by removing redundant descriptors and filtering important descriptors for modeling. The removal of the descriptors will lose important materials information, making it extremely challenging to back-propagate from the descriptors to the materials encoding. The design of the descriptors is easy by forward computation but difficult by the inverse computation. For example, for doped ABO_3-type perovskites, the description of the A-site elements can be described by a weighted average of the atomic properties of all elements in the A-site, but backpropagation of the A-site elements and their doping ratios for a given value could be a very complex task.

In fact, the VSG techniques can be divided into two groups according to the purposes of the generated virtual samples: model construction and model application. The VSG techniques for model construction mainly use sampling or data distribution to expand the training data size to improve the prediction accuracy of the model, including Bootstrap, Monte Carlo, PSO, MTD, GMM, RF, and GAN. The VSG techniques for model application mainly use machine learning models to discover or design virtual samples from a wide materials space that satisfy the target properties to reduce experimental costs for experimenters, including search algorithms for materials discovery and inverse design algorithms for materials design.

CONCLUSION AND OUTLOOKS

This review discusses the applications of VSG techniques in materials science in the context of cutting-edge research achievements. This review summarizes the commonly used VSG algorithms for data expansion, searching algorithms for materials discovery, and materials inverse design algorithms in materials design and discovery and briefly introduces the research cases of virtual samples in materials design and discovery in recent years. VSG is a very promising technology with both opportunities and challenges. Here, we propose the following possible future directions in materials science:

(1) Theoretical support and algorithmic details: As mentioned above, the VSG techniques for expanding the sample size of the training set have been successful in applications, but their theoretical support needs to be improved. For small data with very limited sample size, the improvement of model accuracy by increasing the data size to model is a probable event, but the principle and algorithmic details of VSG still need to be highlighted. For example, GMM assumes that all data points are generated from a mixture of a finite number of Gaussian distributions. However, further hypothesis testing is necessary to determine the consistency of the true data distribution with the assumption of a Gaussian distribution. The study cases in the publications mentioned above do not evaluate different VSG techniques with the same data. To evaluate the performance of the algorithm, it is essential to assess it using a standard benchmark dataset before applying it to the material dataset. However, considering that the material dataset can exhibit significant variations depending on the specific material system and its corresponding encoding, it is crucial to select the optimal method after evaluating multiple VSG techniques. Furthermore, it is vital to validate the validity of the virtual samples, which involves confirming their existence and ensuring the reasonableness of their descriptor values. This verification process requires conducting experiments and calculations to evaluate the generated virtual samples. Rationality and adaptability could be adopted as evaluation criteria for VSG

techniques. Rationality assesses the proximity of generated virtual samples to the real data space, while adaptability measures the generalizability of the VSG method across different domains. Additionally, the investigation of the optimal number of virtual samples generated represents a future direction for development. Insufficient generation of virtual samples may lead to a lack of information, which cannot adequately compensate for small sample sizes, thus limiting the generalization ability. Conversely, an excessive number of virtual samples can disrupt the information of the original data, and errors within the virtual samples may negatively impact the prediction accuracy of the model.

(2) Inclusiveness to material systems: The VSG techniques for expanding the sample size could generate virtual samples from the data perspective, independent of the complexity of the material system. However, materials searching algorithms and materials inverse design algorithms are not very inclusive to the material systems because of the involved materials encoding. For example, organic materials can be encoded using 2D chemical formulae for SMILES or after structure optimization using 3D information, but the structure optimization of organic materials is a tedious process. Performing structure optimization for all samples in the constructed materials space can indeed impose a significant computational burden. Similarly, inorganic crystalline materials can be encoded using only the components or by adding crystal structure information. Currently, the searching algorithms and inverse design algorithms are successfully applied in materials discovery and design. The materials have all adopted simpler encoding methods, such as the combination of doping elements and ratios, SMILES encoding in 2D, *etc.* Bayesian molecular design algorithms are frequently employed for the inverse design of polymers, as the repeating units in polymers can be viewed as combinations of organic fragments. However, the application of these algorithms in the context of inorganic materials is less common. The choice of materials systems plays a pivotal role in determining the usefulness and effectiveness of such algorithms. In future research, the algorithms can be improved to have stronger compatibility for materials encoding and be able to use different encodings to improve prediction accuracy and search accuracy. Alternatively, the gap between organic and inorganic materials can be bridged through the utilization of transfer learning techniques.

(3) Experimental validation of virtual samples: Virtual samples are samples without any experimental validation, so the experimental validation of virtual samples is very important. Experimental validation not only evaluates the value of the algorithm but also breaks the barrier between virtual and real scenarios, enabling the evaluation of the practical application value of the virtual samples. The experimentally validated virtual samples can also be returned to the training set through the active learning framework to achieve a two-way virtuous cycle of machine learning model and material performance optimization through continuous iteration. The experimental validation of virtual samples still faces certain challenges, including the evaluation of the feasibility and the correspondence between descriptors and chemical formulas. To assess the feasibility of organic molecules, the synthetic accessibility score (SA score) can be employed. By utilizing SA scores, researchers can gauge the practicality of synthesizing organic molecules. In materials searching, researchers often rely on directly searching for the chemical formula of the target materials in order to design experimental protocols. However, in scenarios involving data expansion and inverse design, as previously discussed, inferring the chemical formula of the target sample from descriptor values becomes challenging, especially when the descriptors represent the material components. Experiments can be conducted by generating a substantial number of virtual chemical formulas and subsequently calculating the Euclidean distance between the descriptors of these virtual formulas and the target sample. This approach allows for the identification of the closest chemical formula to the target sample, aiding in the exploration and selection of potential candidates for further investigation and experimental validation.

(4) Statistical analysis of the virtual samples: In materials machine learning, researchers often tend to perform statistical analysis of the modeling dataset to explore patterns to guide materials design and discovery. However, when the sample size of the dataset is very limited, the explored patterns need to be used with especial caution, as the patterns are more applicable to the modeling dataset. After reasonable virtual samples are generated, more generalized patterns can be obtained by statistical analysis of the virtual samples. In addition, the effect of a specific feature on the target variable can be visualized by constructing reasonable virtual samples. For example, sensitivity analysis is performed by constructing reasonable virtual samples with fixed values of other features and set steps of the feature values to be explored, combined with model prediction to visualize the impact of the feature. In our work on the design of high bonding strength YSZ materials using GMM in combination with SVR, sensitivity analysis was used to explore the specific effect of APS parameters on bonding strength and to match the domain knowledge. The patterns derived from the statistical analysis of virtual samples tend to exhibit greater generalizability compared to the original small dataset. However, it is crucial to acknowledge that the plausibility of the virtual samples will significantly impact the statistically obtained patterns. Additionally, it is necessary to validate the patterns identified through statistical analysis with experimental samples.

DECLARATIONS

Authors' contributions

Collected publications and completed the framework of the manuscript: Xu P
Revised the manuscript: Ji X, Li M, Lu W

Availability of data and materials

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

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China (No. 52102140), Shanghai Pujiang Program (No. 21PJD024), the Major Science and Technology Projects of Yunnan Precious Metals Laboratory Co., Ltd (No. YPML-2023050205), the Science and Technology Plan Project of Yunnan Precious Metals Laboratory Co., Ltd (No. YPML-2023050208), and the Open Project of Yunnan Precious Metals Laboratory Co., Ltd (No. YPML-2023050280).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

The Author(s) 2023.

REFERENCES

1. Benton WC. Machine learning systems and intelligent applications. *IEEE Softw* 2020;37:43-9. DOI
2. Zhang X, Jiang Y. Research and application of machine learning in automatic program generation. *Chin j electron* 2020;29:1001-15. DOI
3. Zhang X, Liu C, Suen CY. Towards robust pattern recognition: a review. *Proc IEEE* 2020;108:894-922. DOI
4. Rani S, Lakhwani K, Kumar S. Three dimensional objects recognition & pattern recognition technique; related challenges: a review.

Multimed Tools Appl 2022;81:17303-46. DOI
5. Cipriano LE. Evaluating the impact and potential impact of machine learning on medical decision making. *Med Decis Making* 2023;43:147-9. DOI PubMed PMC
6. Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY. Explainable machine learning in materials science. *npj Comput Mater* 2022:8. DOI
7. Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. *Nanoscale Adv* 2020;2:3115-30. DOI PubMed PMC
8. Wei J, Chu X, Sun X, et al. Machine learning in materials science. *InfoMat* 2019;1:338-58. DOI
9. Fu Z, Liu W, Huang C, Mei T. A review of performance prediction based on machine learning in materials science. *Nanomaterials* 2022;12:2957. DOI PubMed PMC
10. Mohtasham Moein M, Saradar A, Rahmati K, et al. Predictive models for concrete properties using machine learning and deep learning approaches: a review. *J Build Eng* 2023;63:105444. DOI
11. Vivanco-benavides LE, Martínez-gonzález CL, Mercado-zúñiga C, Torres-torres C. Machine learning and materials informatics approaches in the analysis of physical properties of carbon nanotubes: a review. *Comput Mater Sci* 2022;201:110939. DOI
12. Park S, Han H, Kim H, Choi S. Machine learning applications for chemical reactions. *Chem Asian J* 2022;17:e202200203. DOI PubMed PMC
13. Bartel CJ, Trewartha A, Wang Q, Dunn A, Jain A, Ceder G. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput Mater* 2020;6:97. DOI
14. Yu Z, Liu Q, Szlufarska I, Wang B. Structural signatures for thermodynamic stability in vitreous silica: Insight from machine learning and molecular dynamics simulations. *Phys Rev Materials* 2021;5:015602. DOI
15. Yang Z, Gao W. Applications of machine learning in alloy catalysts: rational selection and future development of descriptors. *Adv Sci* 2022;9:e2106043. DOI PubMed PMC
16. Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. *npj Comput Mater* 2021;7:23. DOI
17. Timkina YA, Tuchin VS, Litvin AP, Ushakova EV, Rogach AL. Ytterbium-doped lead-halide perovskite nanocrystals: synthesis, near-infrared emission, and open-source machine learning model for prediction of optical properties. *Nanomaterials* 2023;13:744. DOI PubMed PMC
18. Xu P, Chen H, Li M, Lu W. New opportunity: machine learning for polymer materials design and discovery. *Advcd Theory and Sims* 2022;5:2100565. DOI
19. Martin TB, Audus DJ. Emerging trends in machine learning: a polymer perspective. *ACS Polym Au* 2023;3:239-58. DOI PubMed PMC
20. Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. *npj Comput Mater* 2023;9:42. DOI
21. Swain MC, Cole JM. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model* 2016;56:1894-904. DOI PubMed
22. Li Z, Zhang Z, Xiong B, et al. Materials science database in material research and development: recent applications and prospects. *Frontiers Data Comput* 2020;2:78-90. DOI
23. Stein HS, Sanin A, Rahmanian F, et al. From materials discovery to system optimization by integrating combinatorial electrochemistry and data science. *Curr Opin Electrochem* 2022;35:101053. DOI
24. Schleder GR, Padilha ACM, Acosta CM, Costa M, Fazzio A. From DFT to machine learning: recent approaches to materials science - a review. *J Phys Mater* 2019;2:032001. DOI
25. Lin GSS, Tan WW, Tan HJ, Khoo CW, Afrashtehfar KI. Innovative pedagogical strategies in health professions education: active learning in dental materials science. *Int J Environ Res Public Health* 2023;20:2041. DOI PubMed PMC
26. Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clin Chim Acta* 2005;359:1-26. DOI PubMed
27. Zhu Q, Gong H, Xu Y, et al. A bootstrap based virtual sample generation method for improving the accuracy of modeling complex chemical processes using small datasets. In: IEEE 6th Data Driven Control and Learning Systems Conference; 2017 May 26-27; Chongqing, China. IEEE; 2017. p. 84-88. DOI
28. Han P, Yao X, Zhan J, et al. A bootstrap-bayesian dynamic modification model based on small sample target features. In: Global Oceans 2020: Singapore - U.S. Gulf Coast;2020 Oct 5-30; Biloxi, MS, USA. IEEE; 2020; p. 1-6. DOI
29. Rubin DB. The bayesian bootstrap. *Annal Statist* 1981; 9:130-134. DOI
30. Raeside DE. Monte Carlo principles and applications. *Phys Med Biol* 1976;21:181-97. DOI
31. Gong H, Chen Z, Zhu Q, He Y. A Monte Carlo and PSO based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: an empirical study of petrochemical industries. *Appl Energy* 2017;197:405-15. DOI
32. Valle Y, Venayagamoorthy G, Mohagheghi S, Hernandez J, Harley R. Particle swarm optimization: basic concepts, variants and applications in power systems. *IEEE Trans Evol Computat* 2008;12:171-95. DOI
33. Chen Z, Zhu B, He Y, Yu L. A PSO based virtual sample generation method for small sample sets: applications to regression datasets. *Eng Appl Artif Intell* 2017;59:236-43. DOI
34. Yu L, Zhang X. Can small sample dataset be used for efficient internet loan credit risk assessment? Evidence from online peer to peer lending. *Fin Res Lett* 2021;38:101521. DOI
35. Wu S, Wang B, Zhao J, Zhao M, Zhong K, Guo Y. Virtual sample generation and ensemble learning based image source identification with small training samples. *Int J Digit Crime Forensics* 2021;13:34-46. DOI

36. Li D, Wu C, Tsai T, Lina Y. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Comput Oper Res* 2007;34:966-82. DOI
37. Guo Z, Tang J, Qiao J. An improved virtual sample generation technology based on mega trend diffusion. In: 2019 Chinese Automation Congress (CAC); 2019 Nov 22-24; Hangzhou, China. IEEE; 2020. p. 22-24. DOI
38. Zhu B, Yu L, Geng Z. Cost estimation method based on parallel Monte Carlo simulation and market investigation for engineering construction project. *Cluster Comput* 2016;19:1293-308. DOI
39. Yu X, He Y, Xu Y, Zhu Q. A Mega-Trend-Diffusion and Monte Carlo based virtual sample generation method for small sample size problem. *J Phys Conf Ser* 2019;1325:012079. DOI
40. Shen L, Qian Q. A virtual sample generation algorithm supporting machine learning with a small-sample dataset: a case study for rubber materials. *Comput Mater Sci* 2022;211:111475. DOI
41. Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted gaussian mixture models. *Digit Signal Process* 2000;10:19-41. DOI
42. Xu P, Chen C, Chen S, Lu W, Qian Q, Zeng Y. Machine learning-assisted design of yttria-stabilized zirconia thermal barrier coatings with high bonding strength. *ACS Omega* 2022;7:21052-61. DOI PubMed PMC
43. Talekar B. A Detailed review on decision tree and random forest. *Biosci Biotech Res Comm* 2020;13:245-8. DOI
44. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. *Brief Bioinform* 2023;24:bbad002. DOI PubMed PMC
45. He YL, Hua Q, Zhu QX, Lu S. Enhanced virtual sample generation based on manifold features: applications to developing soft sensor using small data. *ISA Trans* 2022;126:398-406. DOI PubMed
46. Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng* 2023;35:3313-32. DOI
47. Cheng J, Yang Y, Tang X, et al. Generative adversarial networks: a literature review. *KSII T Internet Info* 2020;14:4625-4647. DOI
48. Cui C, Tang J, Xia H, Qiao J, Yu W. Virtual sample generation method based on generative adversarial fuzzy neural network. *Neural Comput Appl* 2023;35:6979-7001. DOI
49. He Y, Li X, Ma J, Lu S, Zhu Q. A novel virtual sample generation method based on a modified conditional Wasserstein GAN to address the small sample size problem in soft sensing. *J Process Control* 2022;113:18-28. DOI
50. Zhu Q, Hou K, Chen Z, Gao Z, Xu Y, He Y. Novel virtual sample generation using conditional GAN for developing soft sensor with small data. *Eng Appl Artif Intell* 2021;106:104497. DOI
51. Aggarwal S, Singh P. Cuckoo, Bat and Krill Herd based k-means++ clustering algorithms. *Cluster Comput* 2019;22:14169-80. DOI
52. Liu Z, Guo J, Chen Z, et al. Swarm intelligence for new materials. *Comput Mater Sci* 2022;214:111699. DOI
53. Yan S, Wang Y, Gao Z, Long Y, Ren J. Directional design of materials based on multi-objective optimization: a case study of two-dimensional thermoelectric SnSe. *Chinese Phys Lett* 2021;38:027301. DOI
54. Shim S, Park WB, Han J, et al. Optimal composition of Li argyrodite with harmonious conductivity and chemical/electrochemical stability: fine-tuned via tandem particle swarm optimization. *Adv Sci* 2022;9:e2201648. DOI PubMed PMC
55. Zheng R, Zhang C. Optimized design of absorbing structural materials using a particle swarm optimization algorithm. *Mod Def Technol* 2019;47:88-93. (in Chinese) Available from: https://xueshu.baidu.com/usercenter/paper/show?paperid=1k2b08n0gs0d0ep0b03q0040ef231476&site=xueshu_se. [Last accessed on 6 Jul 2023].
56. Chen L, Liao C, Lin W, Chang L, Zhong X. Hybrid-surrogate-model-based efficient global optimization for high-dimensional antenna design. *PIER* 2012;124:85-100. DOI
57. Xu B, Cai Y. A multiple-data-based efficient global optimization algorithm and its parallel implementation for automotive body design. *Adv Mech Eng* 2018;10:168781401879434. DOI
58. Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. *Comput Chem Eng* 2018;108:250-67. DOI
59. Raponi E, Fiumarella D, Boria S, Scattina A, Belingardi G. Methodology for parameter identification on a thermoplastic composite crash absorber by the sequential response surface method and efficient global optimization. *Compos Struct* 2021;278:114646. DOI
60. Zhao W, Zheng C, Xiao B, et al. Composition refinement of 6061 aluminum alloy using active machine learning model based on Bayesian optimization sampling. *Acta Metall Sin* 2021;57:797-810. DOI
61. Xue D, Balachandran PV, Hogden J, Theiler J, Xue D, Lookman T. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun* 2016;7:11241. DOI PubMed PMC
62. Zhang Q, Hwang Y. Sequential model-based optimization for continuous inputs with finite decision space. *Technometrics* 2020;62:486-98. DOI
63. Li B, Ma JY, Hu K, et al. A method for parameter identification of distribution network equipment based on sequential model-based optimization. *Int Trans Electr* 2022;2022:1-12. DOI
64. 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* 2022;7:21583-94. DOI PubMed PMC
65. Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. *Multimed Tools Appl* 2021;80:8091-126. DOI PubMed PMC
66. Leardi R. Genetic algorithms in chemistry. *J Chromatogr A* 2007;1158:226-33. DOI
67. Lim Y, Park J, Lee S, Kim J. Finely tuned inverse design of metal–organic frameworks with user-desired Xe/Kr selectivity. *J Mater Chem A* 2021;9:21175-83. DOI

68. Dong R, Dan Y, Li X, Hu J. Inverse design of composite metal oxide optical materials based on deep transfer learning and global optimization. *Comput Mater Sci* 2021;188:110166. DOI
69. Toropov AA, Rasulev BF, Leszczynska D, Leszczynski J. Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents. *Chem Phys Lett* 2008;457:332-6. DOI
70. Wang Y, Zeng Q, Wang J, Li Y, Fang D. Inverse design of shell-based mechanical metamaterial with customized loading curves based on machine learning and genetic algorithm. *Comput Methods Appl Mech Eng* 2022;401:115571. DOI
71. Maurizi M, Gao C, Berto F. Inverse design of truss lattice materials with superior buckling resistance. *npj Comput Mater* 2022;8:247. DOI
72. Nigam A, Pollice R, Aspuru-Guzik A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digit Discov* 2022;1:390-404. DOI PubMed PMC
73. Greenhill S, Rana S, Gupta S, Vellanki P, Venkatesh S. Bayesian optimization for adaptive experimental design: a review. *IEEE Access* 2020;8:13937-48. DOI
74. Jiang M, Chen YM. Survey on Bayesian optimization algorithm. *Comput Eng Des* 2010;31:3254-3259. Available from: https://xueshu.baidu.com/usercenter/paper/show?paperid=ce7eea962163345bf08f16cdc1a3db8b&site=xueshu_se. [Last accessed on 6 Jul 2023]
75. Wu S, Lambard G, Liu C, Yamada H, Yoshida R. iQSPR in XenonPy: a Bayesian molecular design algorithm. *Mol Inform* 2020;39:e1900107. DOI PubMed PMC
76. Wu S, Kondo Y, Kakimoto M, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater* 2019;5:66. DOI
77. Serrão R, Oliveira MR, Oliveira L. Theoretical derivation of interval principal component analysis. *Inf Sci* 2023;621:227-47. DOI
78. Hou S, Riley C. Is uncorrelated linear discriminant analysis really a new method? *Chemom Intell Lab Syst* 2015;142:49-53. DOI
79. Yang C, Ren C, Jia Y, Wang G, Li M, Lu W. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Mater* 2022;222:117431. DOI
80. Wang X, Xu P, Lu T, et al. Inverse design of ternary gold alloy materials with low resistivity. *Mater Chin* 2021;40:251-256. (in Chinese) Available from: https://d.wanfangdata.com.cn/periodical/zgcljz202104002. [Last accessed on 6 Jul 2023]
81. Liu Y, Zou X, Yang Z, et al. Machine learning embedded with materials domain knowledge. *J Chin Ceram Soc* 2022;50:863-76. (in Chinese) Available from: https://www.jccsoc.com/Magazine/Show.aspx?ID=51304. [Last accessed on 6 Jul 2023]