

Review

Open Access



# Multi-objective optimization in machine learning assisted materials design and discovery

Pengcheng Xu<sup>1,\*</sup>, Yingying Ma<sup>2,3,#</sup>, Wencong Lu<sup>4</sup> , Minjie Li<sup>4,\*</sup>, Wenyue Zhao<sup>2,\*</sup>, Zhilong Dai<sup>1,\*</sup>

<sup>1</sup>Hubei Institute of Aerospace Chemotechnology, Xiangyang 441003, Hubei, China.

<sup>2</sup>Tianmushan Laboratory, Beihang University, Hangzhou 311115, Zhejiang, China.

<sup>3</sup>Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, Zhejiang, China.

<sup>4</sup>Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China.

<sup>#</sup>Authors contributed equally.

\***Correspondence to:** Prof. Minjie Li, Department of Chemistry, College of Sciences, Shanghai University, 99 Shangda Road, Shanghai 200444, China. E-mail: minjieli@shu.edu.cn; Prof. Wenyue Zhao, Tianmushan Laboratory, Beihang University, 166 Shuanghong Bridge Street, Pingyao Town, Yuhang District, Hangzhou 311115, China. E-mail: wendyzhao56@buaa.edu.cn; Dr. Zhilong Dai, Hubei Institute of Aerospace Chemotechnology, 58 Qinghe Road, Xiangyang 441003, Hubei, China. E-mail: daizhilong@casc42.cn

**How to cite this article:** Xu, P.; Ma, Y.; Lu, W.; Li, M.; Zhao, W.; Dai, Z. Multi-objective optimization in machine learning assisted materials design and discovery. *J. Mater. Inf.* 2025, 5, 26. <https://dx.doi.org/10.20517/jmi.2024.108>

**Received:** 31 Dec 2024 **First Decision:** 6 Feb 2025 **Revised:** 20 Feb 2025 **Accepted:** 26 Feb 2025 **Published:** 24 Mar 2025

**Academic Editor:** Sergei Manzhos **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

## Abstract

Over the past decades, machine learning has kept playing an important role in materials design and discovery. In practical applications, materials usually need to fulfill the requirements of multiple target properties. Therefore, multi-objective optimization of materials based on machine learning has become one of the most promising directions. This review aims to provide a detailed discussion on machine learning-assisted multi-objective optimization in materials design and discovery combined with the recent research progress. First, we briefly introduce the workflow of materials machine learning. Then, the Pareto fronts in multi-objective optimization and the corresponding algorithms are summarized. Next, multi-objective optimization strategies are demonstrated, including Pareto front-based strategy, scalarization function, and constraint method. Subsequently, the research progress of multi-objective optimization in materials machine learning is summarized and different Pareto front-based strategies are discussed. Finally, we propose future directions for machine learning-based multi-objective optimization of materials.

**Keywords:** Machine learning, multi-objective optimization, materials design, Pareto front



© The Author(s) 2025. **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

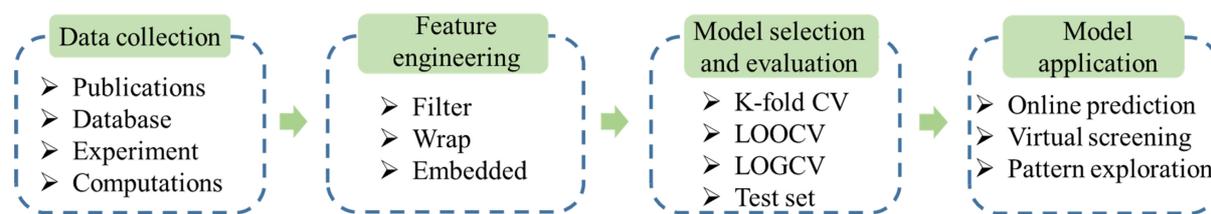
Machine learning, as an interdisciplinary subject covering mathematics, statistics, computer science, and other disciplines, aims to improve the performance of systems<sup>[1-3]</sup>. The core of machine learning includes data and algorithms. By learning patterns from the given data with algorithms, the constructed model can accurately predict the performance of unknown samples. Currently, machine learning has achieved successful applications in various fields<sup>[4-9]</sup>. In materials science, with the development of data storage, materials synthesis and characterization techniques, a vast amount of materials data has been accumulated, offering opportunities for the application of machine learning<sup>[10-13]</sup>. Especially with the introduction of materials genome initiative (MGI), the research on using machine learning to accelerate materials development has experienced rapid growth in the past decade<sup>[14,15]</sup>. The key application of machine learning in materials science is to construct models with high precision for the predictions of unknown materials properties. Machine learning has already been applied to organic molecules, perovskites, alloys, polymers and many other materials systems, achieving numerous successes in performance prediction, materials exploration, compositional optimization, and inverse design<sup>[16-24]</sup>. These advancements indicate that machine learning has become one of the essential tools in the development of materials science.

In practical applications, materials often need to satisfy multiple property constraints, such as strength and ductility in alloys or the catalytic activity, selectivity, and stability in catalysts<sup>[25]</sup>. Optimization of multiple properties belongs to the category of multi-objective optimization<sup>[26,27]</sup>. In materials science, complex relationships between different properties have posed a great challenge to multi-objective optimization. Especially when multiple properties are in conflict with each other, enhancing one property might lead to a decrease in another. In this situation, it could reasonably be assumed that there is some kind of resource within the materials allocated to each other among the multiple properties. Therefore, the Pareto optimization theory in economics can be referred to explore the Pareto front among multiple objectives. The exploration of the Pareto front requires a large number of sample points, which imposes excessive costs through experimentation or first-principles calculations. Due to the excellent prediction and generalization ability, machine learning models combined with heuristic algorithms can calculate the Pareto fronts quickly and accurately<sup>[28,29]</sup>. When the relationship between multiple objectives is unclear, machine learning models can also be constructed separately according to the actual application of materials to transform multi-objective optimization into multiple single-objective optimization tasks. Combined with virtual screening to select materials that meet the performance.

This review aims to provide a detailed discussion on multi-objective optimization in materials machine learning with cutting-edge research progress. In the second section, the workflow of materials machine learning based on multi-objective optimization is introduced. The third section elaborates on the concept of Pareto front and its corresponding computational algorithms. The fourth section discusses commonly used strategies for multi-objective optimization. The fifth section presents the research progress in multi-objective optimization in materials science. The sixth part discusses the impact of different Pareto front-based strategies on the same data based on our previous work. Finally, we propose future development directions for multi-objective optimization in materials machine learning.

## MACHINE LEARNING WORKFLOW FOR MULTI-OBJECTIVE OPTIMIZATION

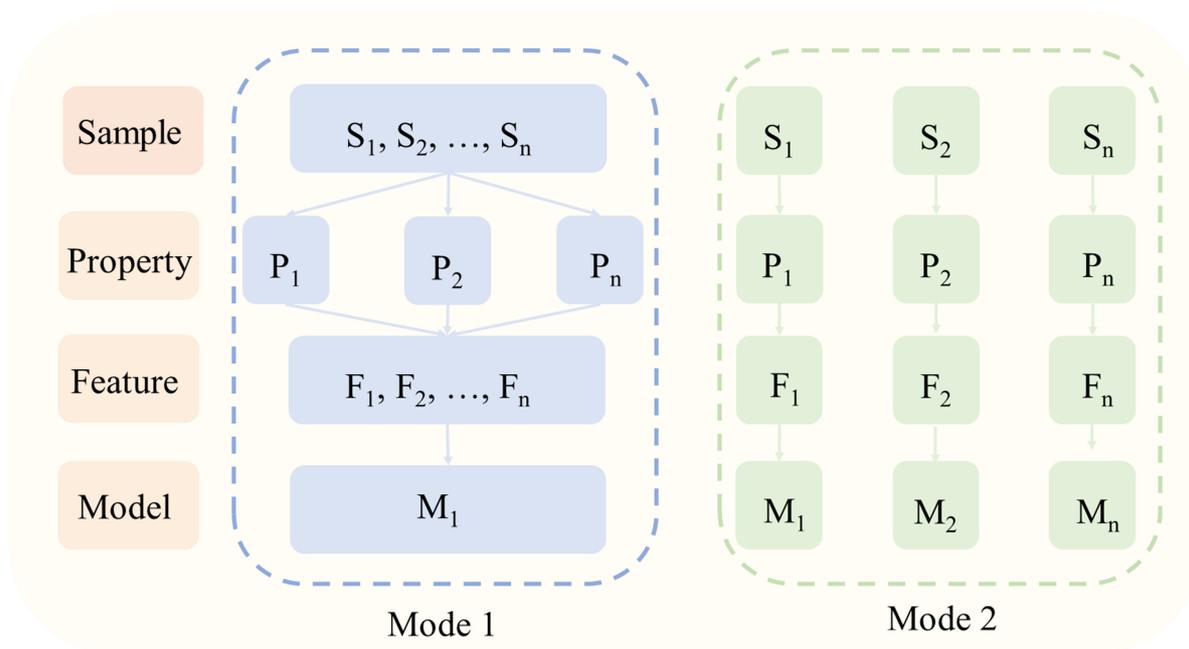
Based on machine learning, the optimization of materials properties, whether single-objective or multi-objective, requires the construction of models. As shown in [Figure 1](#), the workflow can be divided into data collection, feature engineering, model selection and evaluation, and model application<sup>[30-32]</sup>.



**Figure 1.** The workflow of machine learning.

The quality of data usually determines the performance and application value of the models. In the collection of multi-objective data of materials, it is important to consider the correspondence between samples and properties. As shown in Figure 2, two data modes can be commonly adopted. If the data are described in tabular form, with the number of rows as samples and the number of columns as variables (including the target variable and features), mode 1 is just one table, where all the samples have the same features, while mode 2 could be one table for each property, where the sample sizes and features may not be the same. The multi-objective optimization model can be a multi-output model constructed on the same dataset, where each sample has different multiple objectives, and one model can simultaneously predict multiple objectives. Alternatively, the multi-objective optimization model can be transformed into multiple single-objective models, where different objectives have varying samples and features. Multiple models need to be constructed to predict each objective.

The independent variables, also known as features or descriptors, are typically the factors that influence materials properties. Feature engineering refers to encoding the factors into features and selecting the optimal feature subset for modeling. A brief introduction to the common descriptors is shown in Supplementary Table 1, including atomic descriptor, molecular descriptor, crystal descriptor, process parameter descriptor, and domain knowledge descriptor. It should be noted that domain knowledge descriptor refers to extracting descriptors from domain knowledge for modeling, while the current hot direction in materials machine learning is to generate a large number of descriptors from simple mathematical combinations of existing descriptors, and then screen the most core combination of descriptors through machine learning modeling to obtain domain knowledge through the interpretation of descriptors. The most typical method is sure independence screening and sparsifying operator (SISSO), which has been successfully applied as an interpretable machine learning method to correct the tolerance factor of perovskites and propose a principled criterion for strong metal-metal interaction<sup>[33,34]</sup>. For the same materials, different descriptors may result in varying model performance. Materials encoding often generates a large number of features, which may contain irrelevant or redundant information<sup>[35]</sup>. Feature selection aims to identify the features that contribute to model predictions while removing redundant or noisy features<sup>[36-38]</sup>. Based on the relationship between feature selection algorithms and modeling algorithms, the common feature selection methods can be classified into filter, wrapper, and embedded methods<sup>[36,39]</sup>. In specific cases, researchers also tend to use multiple feature selection methods in a stepwise strategy or an ensemble feature selection strategy to select the best feature subset<sup>[40,41]</sup>. Ordillo *et al.* proposed a feature-centered strategy for the prediction of CO and H adsorption energies of adsorbents for CO<sub>2</sub> reduction reactions<sup>[42]</sup>. After evaluating eight different feature selection algorithms with gradient boosted regression and linear regression models, ten and seven features were screened from an initial 86 features for CO and H adsorption energy prediction, respectively, with test set coefficient of determination ( $R^2$ ) of 0.93 and 0.81, respectively. Wang *et al.* proposed an integrated feature selection method called MIC-SHAP, which combines the SHapley Additive exPlanations (SHAP) method and the maximum information coefficient (MIC) method<sup>[41]</sup>. The effectiveness of MIC-SHAP was also evaluated using datasets collected from publications on solid solution strengthening of high entropy alloys, halide chalcogenide band gaps and



**Figure 2.** Different data modes of the multi-objective optimization.

melting points of low melting point alloys.

In model selection and model evaluation, there are similarities between multi-objective and single-objective optimization. Machine learning models constructed with different algorithms on the same dataset can result in models with different predictive accuracy. For specific datasets, it is often necessary to try out multiple algorithms for modeling, combining model evaluation methods and metrics to select the optimal modeling algorithm<sup>[43,44]</sup>. The commonly used evaluation methods include K-fold cross-validation (K-fold CV), leave-one-out cross-validation (LOOCV), leave-out-group cross-validation (LOGCV) and independent test sets<sup>[45-47]</sup>. In order to avoid the random fluctuations caused by data division, it is also necessary to divide the training set and test set several times to validate the stability of the model. After determining the evaluation method, specific metrics are needed to quantify the model performance. For regression tasks, common evaluation metrics include mean relative error (MRE), root mean squared error (RMSE), correlation coefficient (R) and  $R^2$  between predicted and true values. For classification tasks, common evaluation metrics include classification accuracy, true positive rate (TPR), false positive rate (FPR), recall, and precision (P). In addition to predictive accuracy, complexity and interpretability are important factors in model selection. The complexity of the model refers to the dimension of the features used in modeling. The higher the feature dimension would lead to the greater complexity of the model. In the similar prediction accuracy of different models, a subset of features with lower dimensionality is preferred. Interpretability refers to analyzing the predictive mechanisms or important features of a model to obtain domain knowledge to guide material design. With similar prediction accuracy, a modeling method with higher interpretability is preferred.

For a multi-objective optimization model, the fundamental function is to predict multiple properties of materials. Based on this function, the applications of the model can be classified into three categories: online prediction, virtual screening, and pattern exploration. Online prediction involves deploying the model to a backend server and developing user interactive interfaces. Users only need to input necessary information

on the interface to obtain the prediction results. Virtual screening requires the design of a large number of virtual samples. By combining the predicted properties of the model, target materials can be screened based on specific application requirements. Pattern exploration aims to explore the causal relationship between modeling features and target properties. Through statistical analysis methods, such as sensitivity analysis, SHAP and partial dependence plots (PDP)<sup>[48-50]</sup>, crucial features could be explained to guide experiments. The essence of machine learning is that algorithms could learn patterns from data for the predictions of the properties of potential candidates. However, the patterns are ambiguously helpful for materials research in the black box of algorithms. Pattern exploration could analyze and transmit these patterns to researchers, realizing the process of machine learning to learn from machines.

## ALGORITHMS FOR PARETO FRONT

For multi-objective optimization tasks where the objective functions are conflicting, the core step is to find a set of solutions that achieve optimal outcomes across multiple objective functions to form the Pareto front<sup>[51]</sup>. The Pareto front comprises all non-dominated solutions across the multiple objective functions. Solutions on the Pareto front are superior to other solutions in at least one objective function, while being no worse in the remaining objective functions<sup>[52,53]</sup>. The Pareto front aids the researchers in understanding the trade-offs between different objectives and identifying the best solutions achievable under given constraints. In multi-objective optimization, the Pareto front can be continuous or discontinuous. As illustrated in [Figure 3](#), not every region is feasible when searching for the Pareto front<sup>[54]</sup>. Blank areas without solutions might exist, leading to discontinuities in the task. In materials science, it is theoretically feasible to obtain the Pareto front by successive randomized experiments or by exploring a point on the Pareto front, but the limitations are the high cost of the experiments and the tendency to fall into the local Pareto front. With the advancement of heuristic algorithms, computational methods with machine learning have offered a faster way to obtain the Pareto front. Commonly used heuristic algorithms to calculate the Pareto front include multi-objective genetic algorithm, multi-objective differential evolution, multi-objective simulated annealing, multi-objective particle swarm optimization, and multi-objective ant colony optimization. The principles, advantages, and limitations of these algorithms are available in [Supplementary Text](#) and [Supplementary Table 2](#).

## STRATEGY FOR MULTI-OBJECTIVE OPTIMIZATION

### Pareto front-based strategy

In the conflicting objectives in multi-objective optimization, the essence lies in exploring the Pareto front. For instance, in optimizing the hardness and toughness of alloys, the ultimate goal is to find an alloy with both high hardness and toughness. However, due to the trade-off between hardness and toughness, improving one property inevitably leads to a decrease in the other. Leveraging machine learning and multi-objective optimization allows us to compute the Pareto front for the hardness and toughness. First, after collecting multi-target data, multiple single-target machine learning models can be constructed to achieve high-precision prediction for each property. Then, a large number of virtual samples are constructed based on the materials encoding to form the search space, and the machine learning models are utilized to predict multiple properties of the virtual samples in the search space. Next, the Pareto front of the virtual samples could be calculated by combining the multi-objective optimization algorithm. Finally, the samples are selected from the Pareto front for experiments according to the actual application requirements.

The core of the Pareto front-based multi-objective optimization strategy revolves around predicting the multiple properties of virtual samples and computing the Pareto front. Therefore, the success of this strategy heavily relies on the predictive accuracy of the models. Additionally, constructing a large number of virtual samples for model predictions places significant importance on the materials encoding methods and

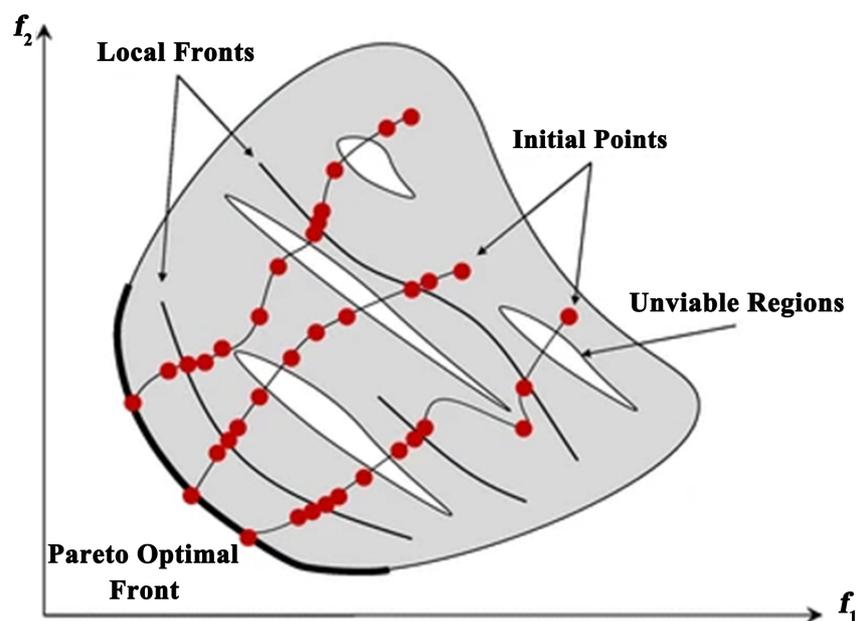


Figure 3. The Pareto front in the multi-objective optimization<sup>[54]</sup>.

descriptors. For instance, when dealing with organic materials, their structures can be encoded into 2D forms, considering only planar structures. Alternatively, encoding them into three-dimensional (3D) forms would consider the impact of 3D configurations. However, optimizing 3D structures to attain the energetically stable state requires substantial effort, adding to the workload post virtual sample generation. While 3D descriptors might yield higher predictive accuracy than 2D ones, they also lead to increased computational efforts during model application. Similarly, inorganic materials can be encoded as combinations of elements and compositions, or with consideration of crystalline parameters for crystal structures. The former would simplify the process of generating virtual samples. Researchers must strike a balance between predictive model accuracy and the complexity of materials encoding. This involves maintaining high predictive accuracy while employing relatively simple encoding rules during virtual sample generation. Furthermore, the Pareto front obtained by multi-objective optimization algorithms exclusively pertains to the Pareto front of virtual sample data, not representing the true Pareto front of the multi-objective properties. This strategy can be combined with active learning, where samples are selected from the Pareto front and experimentally tested, subsequently feeding back into the modeling dataset.

### Scalarization function

In multi-objective optimization, the scalarization function could transform multiple objective functions into a single scalar value, allowing for processing within a single-objective optimization framework<sup>[55,56]</sup>. Multi-objective optimization would involve conflicting objectives, and the role of a scalarization function is to merge these multiple objectives into a single optimization goal, enabling traditional single-objective optimization algorithms to complete the task. After the determination of the optimization objectives, if multiple objective functions have different scales and units, the objective functions need to be normalized to the same scale range. Subsequently, scalarization functions are calculated by assigning weights to each objective function. These weights are used to perform a weighted sum on the normalized objective functions, resulting in a single scalar value - the scalarization function. Next, an appropriate single-objective optimization algorithm is selected to optimize the scalarization function, and parameter adjustments are made to enhance algorithm performance. Once the optimization results are obtained, the solutions derived

from the single-objective optimization need to be reverse-decoded to evaluate their performance and characteristics in the original multi-objective task. In scalarization function-based multi-objective optimization, reverse decoding would be a challenging step. The reverse decoding involves mapping a single optimized solution back to the multi-dimensional, diverse solution space typical of multi-objective tasks. This process might entail nonlinear and intricate mapping relationships. Additionally, complex interrelationships can exist between different objectives in multi-objective tasks. While single-objective optimization might find several relatively optimal solutions, these solutions might not possess the same performance in the multi-objective task. Reverse decoding must address how to map these multiple solutions back to the multi-objective solution space while preserving diversity and maintaining balance. Overcoming these challenges necessitates a combination of domain knowledge, optimization algorithms, mathematical modeling, and experimental validation.

The most common scalarization function is the weighted sum<sup>[57]</sup>, where each objective function is multiplied by a weight and then summed up, expressed as:

$$F(x) = w_1 \times f_1(x) + w_2 \times f_2(x) + \dots + w_n \times f_n(x)$$

Where  $F(x)$  is the scalarized objective function value, and  $w_i$  is the weight of the  $i$ -th original objective function  $f_i(x)$ . The weights in the scalarization function determine the significance of each objective function during the scalarization process. Determining the weights often requires consideration of task nature, domain knowledge, and decision-maker preferences. The main drawback of the weighted sum scalarization function is its potential bias towards objectives with larger weights, neglecting other objectives. This can hinder the attainment of well-balanced solutions in multi-objective tasks. The Chimera scalarization function addresses this issue by not only considering the weights of objective functions but also introducing the concept of tolerance to mitigate potential imbalances arising from the weighted sum scalarization function<sup>[58]</sup>. In the Chimera scalarization function, tolerance is incorporated as an acceptable degradation level relative to the best value for each objective. Tolerance is computed by multiplying the difference between the minimum and maximum values of each objective by a relative tolerance factor, sampled from a uniform distribution in the interval [0.01, 0.5]. The Chimera scalarization function considers both the weights and balance among objective functions to achieve better-balanced solutions for multi-objective tasks.

### Constraint method

The constraint method is commonly used to effectively handle constraints in multi-objective optimization tasks<sup>[59]</sup>. In multi-objective optimization, besides the optimization objectives, there are additional constraints that define the feasible region of the task, which is the set of candidate solutions. The constraint method incorporates these constraints into the objective function of the optimization task, which could also transform the multi-objective optimization task into a single-objective optimization task.

A widely used constraint method is the  $\epsilon$ -constraint<sup>[60]</sup>. The fundamental idea of the  $\epsilon$ -constraint is to introduce a constraint for each optimization objective, restricting its value to be less than or equal to a predefined  $\epsilon$  value. By iteratively treating each optimization objective as the primary objective and the others as constraints, the multi-objective task is decomposed into a series of single-objective tasks with constraints. Subsequently, single-objective optimization algorithms can be employed to solve these tasks, obtaining a series of potential solutions. In materials machine learning, the application of the  $\epsilon$ -constraint method usually corresponds to high-throughput screening. After constructing models for different objectives, a multitude of virtual samples undergo model predictions. According to the practical

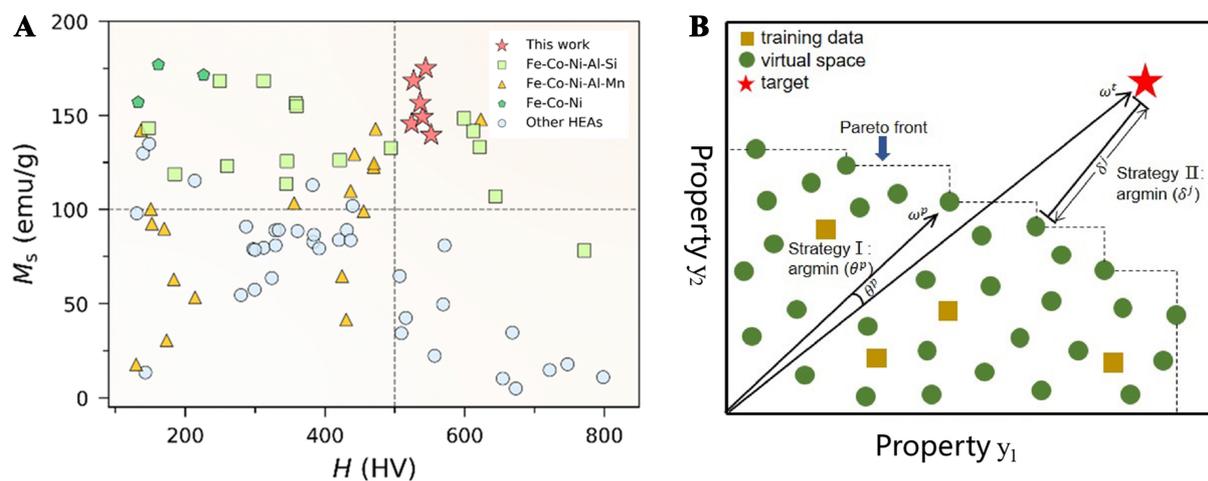
requirements of materials applications, thresholds are set for each objective to conduct tiered screening until the target materials are identified. The  $\epsilon$ -constraint method can also be applied to compute the Pareto front by progressively adjusting  $\epsilon$  values, yielding a series of solutions along the Pareto front. Each  $\epsilon$  value corresponds to a set of solutions on the Pareto front. As  $\epsilon$  values increase, more improvements in objective function values are permitted, consequently obtaining a greater number of non-dominated solutions. Analyzing the solution sets for different  $\epsilon$  values offers insights into various trade-off solutions, aiding in selecting the most suitable solution. However, in the realm of multi-objective optimization in materials science, researchers are inclined to employ heuristic algorithms to compute the Pareto front, while integrating the core concepts of the  $\epsilon$ -constraint method into high-throughput screening. In high-throughput screening, the threshold for each objective is set based on the practical applications, often without the need to consider the Pareto front among multiple objectives. This is commonly utilized for multi-objective optimization scenarios where the objectives are non-conflicting.

## RECENT PROGRESS OF MULTI-OBJECTIVE OPTIMIZATION IN MATERIALS SCIENCE

In this section, the cutting-edge multi-objective research progress in materials machine learning is presented, including the research with and without the Pareto front. The materials systems in the research progress cases include perovskites, alloys, steels, organic small molecules, *etc.*, aiming to provide more inspiration for multi-objective optimization in materials science.

### Research with Pareto front

High-entropy alloys, as functional materials with outstanding magnetic properties, possess vast application potential. However, certain mechanical and magnetic properties of these alloys may be incompatible, such as the contradiction between saturation magnetization and hardness. To address this issue, Li *et al.* employed machine learning combined with multi-objective optimization methods to discover six high-entropy alloys simultaneously exhibiting high saturation magnetization and hardness, which were further experimentally validated<sup>[61]</sup>. Initially, 357 high-entropy alloy samples were collected from published papers, including 139 samples with saturation magnetization, 267 samples with hardness, and 73 samples with both saturation magnetization and hardness. Subsequently, recursive feature elimination was combined with LightGBM, random forest, and XGBoost to rank the feature importance for feature selection. Following a comparison of ten modeling algorithms and parameter optimizations, supporting vector regression (SVR) and LightGBM were identified as the best modeling algorithms for saturation magnetization and hardness, respectively. Then the authors conducted statistical analysis on the samples and constructed a virtual sample space exceeding 1.31 million samples. Bayesian optimization (BO), combined with the expected improvement (EI) method, was utilized to explore high-entropy alloys with simultaneously high saturation magnetization and hardness from the virtual sample space. For candidate discovery, the authors employed three multi-objective optimization strategies:  $\epsilon$ -constraint, weighted sum, and multi-objective evolutionary algorithm. Due to the intervention of EI in BO,  $\epsilon$ -constraint and weighted sum transformed the multi-objective optimization into a single-objective optimization, yielding only one Pareto optimal solution. Adjusting the weights in the weighted sum could affect the final output. On the other hand, the multi-objective evolutionary algorithm of NSGA-II provided a set of Pareto optimal solutions. As shown in [Figure 4A](#), the  $\epsilon$ -constraint, weighted sum, and multi-objective evolutionary algorithm respectively identified 1, 2, and 3 high-entropy alloys with high saturation magnetization and high hardness, all of which were experimentally validated. Ma *et al.* used a framework combining machine learning models and multi-objective optimization algorithms to design high-entropy alloys with high Vickers hardness ( $H$ ) and high compressive fracture strain ( $D$ )<sup>[62]</sup>. First, the  $D$  and  $H$  datasets were constructed, which contained 175 and 467 data points, respectively, after data preprocessing, and 161 features containing atomic parameters, phase parameters, *etc.* were generated for each data based on its alloy composition. Based on the four-step feature selection method, 12 and eight features were selected to model the SVR and LightGBM optimization



**Figure 4.** (A) Experimental saturation magnetization and hardness values for the designed alloying compositions in this work compared with other HEAs in the dataset<sup>[61]</sup>; (B) A schematic of our design strategy for selecting the best trade-off solution<sup>[63]</sup>. HEAs: High-entropy alloys.

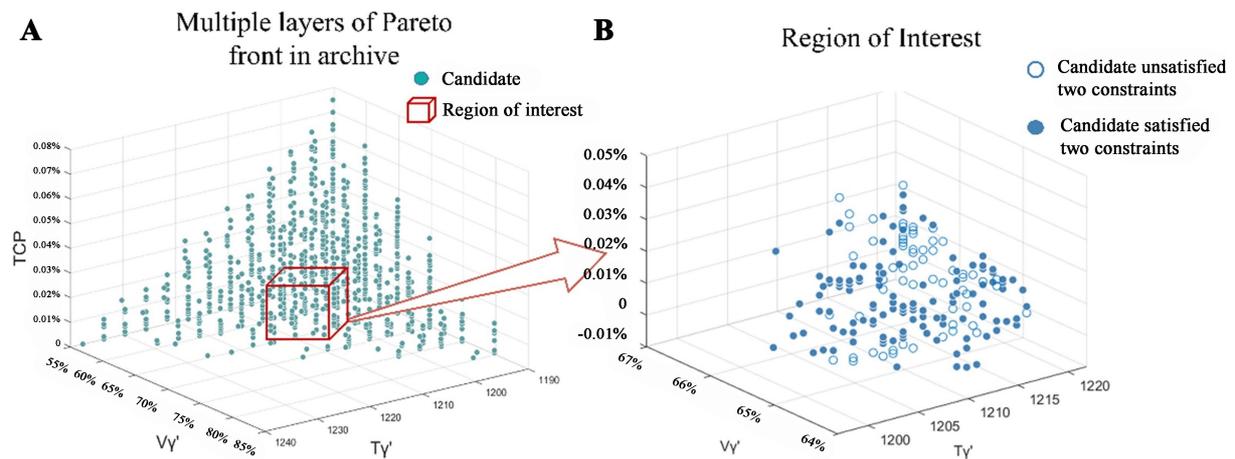
algorithms with  $D$  and  $H$  as target variables, respectively. The  $R^2$  values of the parameter-optimized models were 0.76 and 0.90 in 10-fold cross-validation, respectively. Next, NSGA-II and virtual sample generation were used to search for the optimal alloy compositions, and 105 candidates were obtained as the approximate solutions of the Pareto-optimal solutions. The error analysis revealed that candidate samples containing Mo, Nb, Zr, Hf and Ta may have smaller prediction errors, and four candidate samples with high hardness and appropriate ductility were selected for experimental validation. Three of the candidate samples had an excellent combination of hardness and ductility, with  $H$  values exceeding 600 hypervolume (HV) and  $D$  values exceeding 10%. In addition, the  $D$  values of these candidate samples increased by 135.8%, 282.4%, and 194.1%, respectively, compared to high-entropy alloys with similar hardness levels. These research findings offer valuable guidance for the optimization design and application of high-entropy alloys.

Chen *et al.* combined multi-objective optimization strategy with active learning to iteratively recommend experimental plans, aiming to enhance the yield strength (YS) and fracture strain of cast ZE62 magnesium alloy<sup>[63]</sup>. Firstly, they conducted an orthogonal experimental design, synthesizing and characterizing ten samples of cast ZE62 magnesium alloy to obtain the YS and fracture strain as the target variables. The thermal processing parameters were used as modeling descriptors. SVR with radial basis function kernel was used to construct models. Under various combinations of thermal processing parameters within their respective ranges and step sizes, a total of 23,760 thermal processing schemes were generated. Using machine learning models, the YS and fracture strain for different thermal processing schemes were obtained to calculate the Pareto optimal solutions. With the final optimization objectives set, the research team employed two multi-objective optimization strategies in Figure 4B to obtain experimental plans. Strategy 1 utilized Pareto front analysis, defining two vectors,  $w_i$  and  $w_p$ , representing the origin to optimization objectives and the origin to the Pareto front in the search space, respectively. By minimizing the angle between these two vectors, the optimal experimental plans were obtained. Strategy 2 transformed the bi-objective optimization task into a single-objective optimization task by calculating distances, and then selected the points closest to the optimization objectives as the optimal experimental plans. After conducting experiments with the optimized plans and adding the results to the dataset, the process of modeling, prediction, and Pareto optimal solution search steps was iteratively executed until reaching the optimization objectives. The research results demonstrated that as the number of iterations increased, the

points selected by both strategies gradually approached the target points. In the fourth iteration, strategy 2 recommended the best thermal processing parameters, resulting in a 27% increase in YS and a 13.5% increase in fracture strain for the cast alloy. This strategy proved effective in solving multi-objective optimization tasks in materials composition and processing parameter design.

A combined strategy of machine learning and multi-objective optimization was employed by Deng *et al.* to design a Ni-based high-temperature alloy with high  $\gamma'$  solvus temperature, high  $\gamma'$  volume fraction, and low topologically close-packed (TCP) phase content<sup>[64]</sup>. Firstly, a sample space of  $3.3974 \times 10^{17}$  samples was generated by varying the composition of nine elements in the  $\text{Ni}_a\text{Al}_b\text{Ti}_c\text{Ta}_d\text{Nb}_e\text{Hf}_f\text{Cr}_g\text{Co}_h\text{W}_i\text{Mo}_j\text{C}_k$  system (with fixed C content and Ni as the remaining element). Subsequently, 56,909 samples were obtained through random sampling with their target properties calculated using the Thermo-Calc software. After comparing five modeling algorithms, the backpropagation neural network (BPNN) model demonstrated the best performance in predicting the  $\gamma'$  solvus temperature ( $T_{\gamma'}$ ),  $\gamma'$  volume fraction ( $V_{\gamma'}$ ), and TCP phase content, with  $R^2$  values of 0.9788, 0.9623, and 0.9268, respectively. Based on Pearson correlation coefficients (PCC) and domain knowledge analysis, the optimization objectives for the Ni-based high-temperature alloy were determined as the  $V_{\gamma'}$ ,  $T_{\gamma'}$ , and TCP phase content, close to 65%, 1,210 °C, and 0.01%, respectively. The constraints were set as the liquidus temperature ( $T_l$ ) greater than 40 °C and the processing window ( $T_s - T_{\gamma'}$ ) greater than 1,300 °C. As shown in Figure 5, the NSGA-II algorithm was employed to compute the Pareto-optimal solutions for these three objectives. Two hundred samples met the optimization objectives, in which 133 samples satisfied both the optimization objectives and the constraints. The uniformity of the multi-objective differential evolution (MOEA) predictions and the thermal calculations for these 133 samples demonstrated the reliability and effectiveness of the constructed models for predicting materials properties and optimizing the Ni-based high-temperature alloy. Finally, nine samples were uniformly selected from the 133 samples for experimental validation. The experimental results showed close agreement between the measured values and the model predictions for the three objectives. Moreover, the microstructural parameters of sample  $\text{NiCr}_{6.5}\text{Co}_{7.05}\text{W}_{3.45}\text{Mo}_{2.85}\text{Al}_{5.5}\text{Ti}_{2.5}\text{Ta}_{1.15}\text{Nb}_{1.4}\text{C}_{0.1}$  were found to be very similar to those of the commercial Ni-based high-temperature alloy K424. These research findings provide evidence for the effectiveness and feasibility of the combined strategy of machine learning and multi-objective optimization in designing and optimizing high-performance Ni-based high-temperature alloys.

Aluminum alloys exhibit a high strength-to-weight ratio and sufficient ductility, but their wear resistance is relatively poor. Aluminum-based metal matrix composites can offer higher strength, hardness, and excellent wear resistance. Banerjee *et al.* improved the frictional mechanical properties of alumina-reinforced aluminum-based composites (AMC) using an artificial neural network (ANN) combined with a genetic algorithm<sup>[65]</sup>. Initially, they collected 500 samples from published papers, including pure aluminum, aluminum alloys, and alumina-reinforced AMC. Each sample included information on chemical composition, morphology of alumina particles, process parameters, frictional test parameters, mechanical properties, and wear performance. The optimization goals consisted of four mechanical properties: YS, ultimate tensile strength (UTS), elongation (%*Tel*), and hardness; as well as two wear properties: Wear Rate and Friction Coefficient. Utilizing ANN, all six models were constructed with  $R$  values all above 0.9. Ten multi-objective tasks were constructed, with six properties and their heat-treatment status as constraints. The Pareto optimal solutions were calculated using the genetic algorithm. The analysis of the multi-objective Pareto front revealed that copper, manganese, magnesium, and alumina were crucial variables in designing composites with improved frictional mechanical properties. The preferred copper content varied between 4-4.5 wt%, manganese was around 1 wt%, magnesium ranged from 1-2 wt%, and the alumina content increased with the increase in YS or hardness, with a preferred range of 10-15 wt%. The highlight of this work lies in utilizing Pareto front analysis to provide a wealth of information about variable design.



**Figure 5.** Multi-layer of Pareto front obtained by MOEA. (A) Pareto front of  $V_{\gamma}$ ,  $T_{\gamma}$ , and TCP phases content; (B) Region of Interest<sup>[64]</sup>. MOEA: Multi-objective differential evolution; TCP: topologically close-packed.

In the research of coating materials, the ratio of hardness ( $H$ ) to Young's modulus ( $E$ ) is commonly used as an indicator of coating performance. Materials with a high  $H/E$  ratio can tolerate more deformation, indicating excellent hardness and toughness. Wu *et al.* collected data of high-entropy nitride coatings in both quinary and hexagonal systems and trained machine learning predictive models<sup>[66]</sup>. Using a multi-objective optimization approach, they successfully prepared high-entropy nitride coatings with a combination of high  $H$  and low  $E$ , which were subsequently verified through experiments. Firstly, 301 samples were collected from the published papers, covering high-entropy nitride coatings with quinary and hexagonal elements. Among them, 167 samples were used for the hardness prediction, and 134 samples for the Young's modulus prediction. The modeling features included composition and process parameters, such as the deposition temperature, target power, and bias voltage of the coatings. After an algorithm comparison, the Gaussian process model showed the best performance for hardness and Young's modulus predictions. After parameter optimization using Bayesian algorithms, the  $R^2$  values of the hardness and Young's modulus models in the test set were 0.67 and 0.74, respectively. In the construction of the search space, the dataset only included 44 sets of hexagonal alloy element systems, including (AlCrTiZrMoTa)N, (AlCrSiTiVNb)N, (AlCrSiTiZrTa)N, (AlCrZrMoMnNi)N, and (AlCrSiTiZrMo)N. The research indicated that adding an appropriate amount of Si could optimize the mechanical properties of the coatings by refining grains and altering microstructures. Mo elements easily reacted with N elements to form MoN, which exhibited high hardness, high melting point, and good thermal stability. Based on these factors, a new high-entropy nitride system, (AlCrSiTiMoTa)N, was selected for subsequent multi-objective optimization. After predicting the search space using the model, the NSGA-II was used to calculate the Pareto front of  $H$  and  $E$  in the search space, aiming to find the optimal combination of high  $H$  and low  $E$ . By selecting samples on the Pareto front with lower tangent slopes and higher  $H$  values, the optimal samples were determined and subjected to experimental verification. The experimental results showed that the measured hardness was higher than the predicted results by NSGA-II, with an error of about 6.6%, while the measured modulus was lower than the predicted results, with an error of about 11.5%.

Zhang *et al.* combined a neural network (NN)-based machine learning framework with a genetic algorithm to achieve multi-objective performance optimization of conflicting iron-based metallic glasses<sup>[67]</sup>. Firstly, they collected critical casting sizes ( $D_{\max}$ ), saturation magnetization ( $B_s$ ), and ductility ( $\nu$ ) data of 589, 360, and 21 iron-based metallic glasses from existing works. The input features include alloy compositions (COMP), elemental properties (ELEM), and combinations of alloy compositions and elemental properties

(COMP + ELEM). Then, they used a NN to model the data and explored the impact of hidden layers, exclusion methods, and input features on the prediction accuracy of the NN. Experimental results show that the exclusion method can improve model performance, and due to skewed data distribution, predicting  $D_{\max}$  requires more hidden layers. Moreover, compared to using ELEM or combinations of alloy compositions and ELEM as inputs, using only alloy compositions yields better prediction performance. The prediction models for  $D_{\max}$  and  $B_s$  have  $R^2$  values of 0.874 and 0.963, respectively. Finally, this framework was validated on two systems,  $\text{Fe}_{0.83}\text{C}_1\text{B}_x\text{Si}_y\text{P}_{16-x-y}$  and  $\text{Fe}_x\text{Co}_y\text{Ni}_{72-x-y}\text{B}_{19.2}\text{Si}_{4.8}\text{Nb}_4$  with the genetic algorithm calculating the Pareto front of  $D_{\max}$ ,  $B_s$ , and  $v$  for these two systems. The analysis results are in good agreement with the experimental results, providing insights for the multi-objective optimization of iron-based metallic glasses designed for advanced multifunctional applications.

BO is a supervised learning technique. It assesses a set of initial input points with favorable space-filling characteristics, constructs a surrogate model, and then employs an acquisition function to select the most promising inputs for subsequent evaluations. This process continuously updates the surrogate model until the stopping criterion is met. By applying the multi-objective BO approach, Gao *et al.* have established an active learning framework with the objective of designing 3D printing materials possessing superior hardness, strength, and toughness<sup>[68]</sup>. The study initially standardized the experimental process of 3D printing based on digital light processing (DLP) technology. A total of 24 samples were prepared and their mechanical properties, including hardness, flexural strength, tensile strength, and elongation at break, were evaluated using six distinct base resin materials while ensuring the consistency of printing parameters. Subsequently, a prediction model was developed for the aforementioned four mechanical properties by using the resin components as characteristic variables and Gaussian process regression (GPR) as a surrogate model. The initial design space was formed by combining the six base resin materials with the Latin hypercube sampling (LHS) strategy and 24 initial sample data. To augment the optimization efficiency, a BO algorithm was introduced to iteratively adjust the formulation composition of the 3D printing materials. The noise expectation hypercubic improvement algorithm (qNEHVI) was utilized as the acquisition function to balance the relationship between exploring new samples and exploiting existing information. In each iteration, four optimized formulations were chosen for experimental validation. Additionally, the authors proposed a comprehensive evaluation mechanism based on the optimized properties of the materials, as given in

$$\partial = \sum_{i=1}^4 \frac{|y_i - y_i^{\max}|}{y_i^{\max}}$$

$$\beta = \frac{\sum_{j=1}^4 \partial_j}{4}$$

Where  $y_i$  and  $y_i^{\max}$  are the performance of the current sample and the best sample, respectively.  $\partial$  represents the sum of the relative errors of the current mechanical properties of the four samples with respect to the optimal properties, while  $\beta$  denotes the average  $\partial$  value of the recommended four samples. When the rate of change of  $\beta$  is less than 0.5% for three consecutive times, the iteration is determined to have converged. After 15 iterations, the optimization process reaches a convergent state, and the designed 3D printed material exhibits a remarkable enhancement in mechanical properties: the hardness is elevated from 78.3 to 84.5 HD, the flexural strength is augmented from 53.7 to 89.2 MPa, the tensile strength is increased from 30.4 to 45.6 MPa, and the elongation at break is substantially increased from 4.9% to 13.6%. It is noteworthy that, in contrast to the NSGA-II algorithm commonly employed in traditional multi-objective optimization to identify the optimal solution by calculating the Pareto front, qNEHVI is adopted as the acquisition

function in this study. It not only exhibits good anti-noise performance and low computational complexity but also can more effectively balance the process of exploring and optimizing the materials, thereby significantly enhancing the optimization efficiency.

Morales-Hernández *et al.* proposed two constrained BO acquisition functions for addressing a bi-objective optimization task associated with the material bonding process, specifically minimizing the production cost while maximizing the fracture strength<sup>[69]</sup>. The acquisition function constrained expectancy improvement (cMEI) - stochastic kriging (SK) transforms the multi-objective issue into a single-objective one with the assistance of an augmented Tchebycheff scalarization function. In each BO iteration, it trains a SK surrogate model solely for the scalarized objective and explicitly accounts for the noise of the scalarized objective. A cMEI acquisition function is combined with the probability of feasibility (PoF) for computation. Conversely, the acquisition function cEHVI - SK does not employ the scalarization method. It trains separate SK surrogate models for each objective and selects the filling points by computing the HV improvement, also in conjunction with the PoF to consider the constraints. To evaluate the proposed optimization approach, it was experimentally compared with five state-of-the-art constrained evolutionary multi-objective optimization algorithms (C-NSGA-II, C-MOEA/D, C-TAEA, C-MOPSO, OK-C-NSGA-II, and C-K-RVEA). The experiments were conducted using the Matlab process simulator. Ranges of values for each process parameter were defined, and each algorithm was evaluated in a specific manner for 60 process configurations, replicated five times. The BO algorithm initiated with 20 initial configurations and added one fill point per iteration, whereas the evolutionary algorithm utilized the same initial population and generated a new population. The quality of the Pareto front was assessed using HV and inverted generational distance (IGD) + metrics. HV characterizes the volume dominated by the resulting front in the target space, and IGD + represents the average minimum inverse generation distance to the nearest member of the true Pareto front. Larger HV and lower IGD + values signify higher-quality Pareto fronts. The experimental results demonstrated that the BO algorithm with the cMEI - SK acquisition function exhibited better performance in cost optimization, while the BO algorithm with the cEHVI - SK acquisition function had an edge in fracture strength optimization.

Fang *et al.* proposed a machine learning strategy for the multi-objective optimization of joint strength and extended area of Cu-Ag-Zn-Mn-Ni-Si-B-P brazing materials to achieve a comprehensive enhancement of brazing performance<sup>[70]</sup>. A total of 122 samples of Cu-Ag-Zn-Mn-Ni-Si-B-P brazing materials, along with their experimental values of joint strength and extended area, were obtained via high-throughput experiments. To screen the key component characterization, 37 data points with the same brazing process but different compositions and properties were employed. The variance and mean of the physical and chemical parameters of the elements, along with the process parameters of brazing temperature and brazing holding time, were used for feature construction, amounting to a total of 102 features. After feature selection and model comparison using the correlation coefficient, genetic algorithm, feature weight ranking, and exhaustive enumeration method in combination with SVR, tree regression, integrated tree regression, and GPR, the results indicated that the joint strength and extended area models constructed by six key features and four key features combined with SVR exhibited the best performance. The  $R^2$  values of the training set were 0.94 and 0.76, respectively. Subsequently, the combination of predictive models was employed to simultaneously optimize the brazing performance. The BO strategy, which is based on BO and utilizes the "EI" as the acquisition function, was used to construct the multi-objective acquisition function MOEI. This function was then used to iteratively optimize the alloy compositions and brazing process. Each time, the three combinations with the largest MOEI values were selected for experimental validation, and the data were subsequently added to the training set to continue the optimization. After three iterations, the optimal optimized alloy Cu-18Ag-2.8Zn-22.5Mn-15.5Ni-0.4Si-0.3B-0.3P and its matching brazing process (960 °C,

9 min) were obtained. The experimentally verified joint strength and extended area reached 346 MPa and 387 mm<sup>2</sup>, respectively. The Pareto front optimization strategy based on the NSGA-II uses the machine learning performance prediction model as the objective function. The NSGA-II is applied to rapidly screen the target components and brazing process space, optimize the chromosome through genetic evolution operations, calculate the target value after decoding, and obtain the optimization results by non-dominated sorting and screening. By using this Pareto front optimization strategy of the NSGA-II, the optimal optimized alloy Cu-14Ag-4Zn-24Mn-13.5Ni-0.4Si-0.3B-0.3P and its matched brazing process (945 °C, 12 min) were obtained. The experimentally verified joint strength and extended area reached 356 MPa and 412 mm<sup>2</sup>, respectively. Upon comparing the experimental validation results of the two combined process optimization design strategies, it was found that the NSGA-II genetic algorithm has a more favorable performance improvement effect.

### Research without Pareto front

The ABO<sub>3</sub> perovskites are considered one of the most promising photocatalytic materials. Their photocatalytic properties can be evaluated through multiple objectives. Tao *et al.* utilized a machine learning approach to design a multi-objective stepwise design strategy for discovering high-performance perovskite photocatalysts<sup>[71]</sup>. The workflow is shown in Figure 6. Initially, 170, 172, and 117 ABO<sub>3</sub> perovskites with experimental bandgap ( $E_g$ ), specific surface area (SSA), and grain size (CS) were collected from the publications. The original features included 20 atomic descriptors and three experimental condition parameters. After feature selection and algorithm comparison, gradient boosting regression (GBR) was employed to construct the  $E_g$  and CS models, while SVR was used to construct the SSA model. After model optimization, the average R values of the  $E_g$ , SSA, and CS models through 100 random splits of the LOOCV were 0.9079, 0.8525, and 0.8702, respectively. Virtual screening identified 35 candidates with suitable  $E_g$ , high SSA, and small CS. To test the photocatalytic performance of the proposed perovskite candidates, 80 perovskite samples with experimentally measured hydrogen evolution rates were collected from the publications. A hydrogen evolution rate model was constructed using the experimental condition parameters and the  $E_g$ , SSA, and CS predicted by the above models as features. After model optimization, the R of LOOCV reached 0.9173, indicating an accurate prediction of the hydrogen evolution rate for the candidates. The predicted results showed that the screened candidates exhibited hydrogen evolution rates above 6,000  $\mu\text{mol}\cdot\text{h}^{-1}\cdot\text{g}^{-1}$  in photocatalysis, demonstrating satisfactory performance.

Cai *et al.* proposed a stepwise materials screening framework that combines high-throughput computing and machine learning techniques for the rapid discovery of multi-objective excellent lead-free hybrid organic-inorganic double perovskite (HOIDP) materials for solar cells<sup>[72]</sup>. The process of discovering novel HOIDPs according to the combination of machine learning and density functional theory (DFT) calculation for photovoltaic application is shown in Figure 7. The study focused on three optimization objectives: formation energy, bandgap, and Debye temperature. Initially, 4,456 HOIDP samples were collected and their formation energies were calculated using DFT. Among them, 425 samples exhibited direct bandgaps, which were further validated using Perdew-Burke-Ernzerhof (PBE)-DFT calculations, and a PBE bandgap model was constructed. Considering that the PBE method often underestimates semiconductor bandgap, the research team also constructed a Heyd-Scuseria-Ernzerhof (HSE) bandgap model based on 663 compounds with a precise HSE bandgap. Subsequently, the 425 samples with direct bandgap were used to compute the Debye temperatures using DFT for building the Debye temperature prediction model. After comparing seven algorithms, the GBR model performed optimally in predicting performance, bandgap, and Debye temperature, with R<sup>2</sup> values of 0.990, 0.920, 0.870, and 0.990, respectively. The research team adopted a stepwise screening strategy to select the most promising candidates. Firstly, they combined 32 A-site organic cations, 58 B-site cations, and four X-site halogens to construct 180,038 electrically neutral HOIDP virtual samples. Subsequently, using tolerance factors and octahedral factors, they filtered down to 25,983

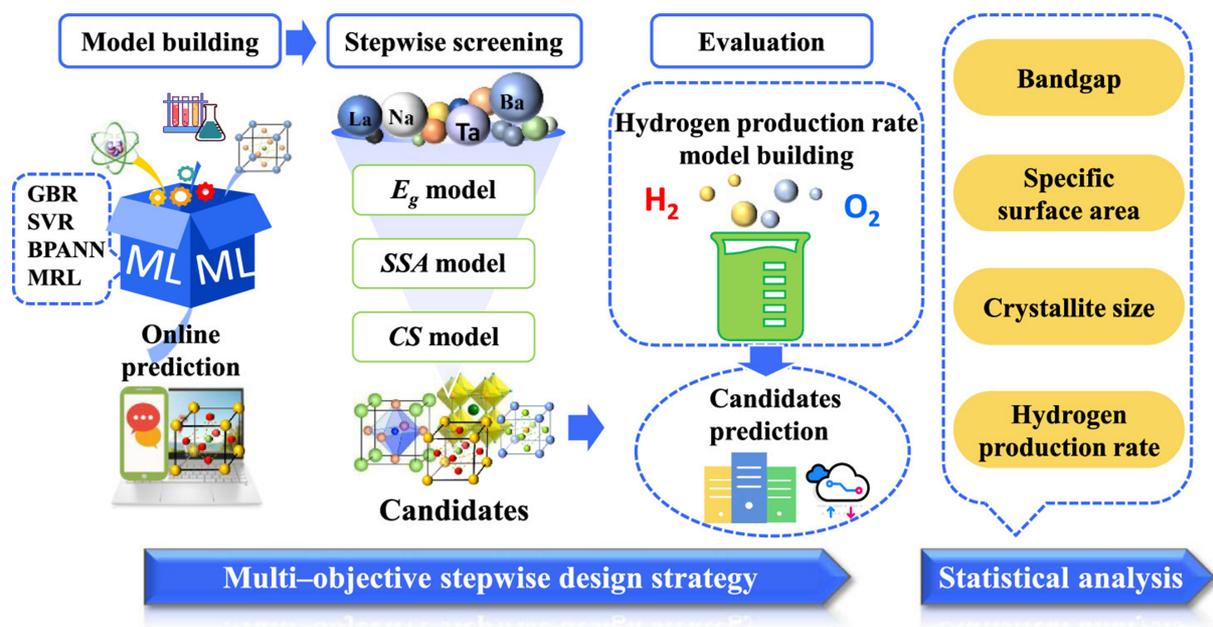


Figure 6. Workflow of multi-objective stepwise design strategy-assisted design of high-performance perovskite oxide photocatalysts [71].

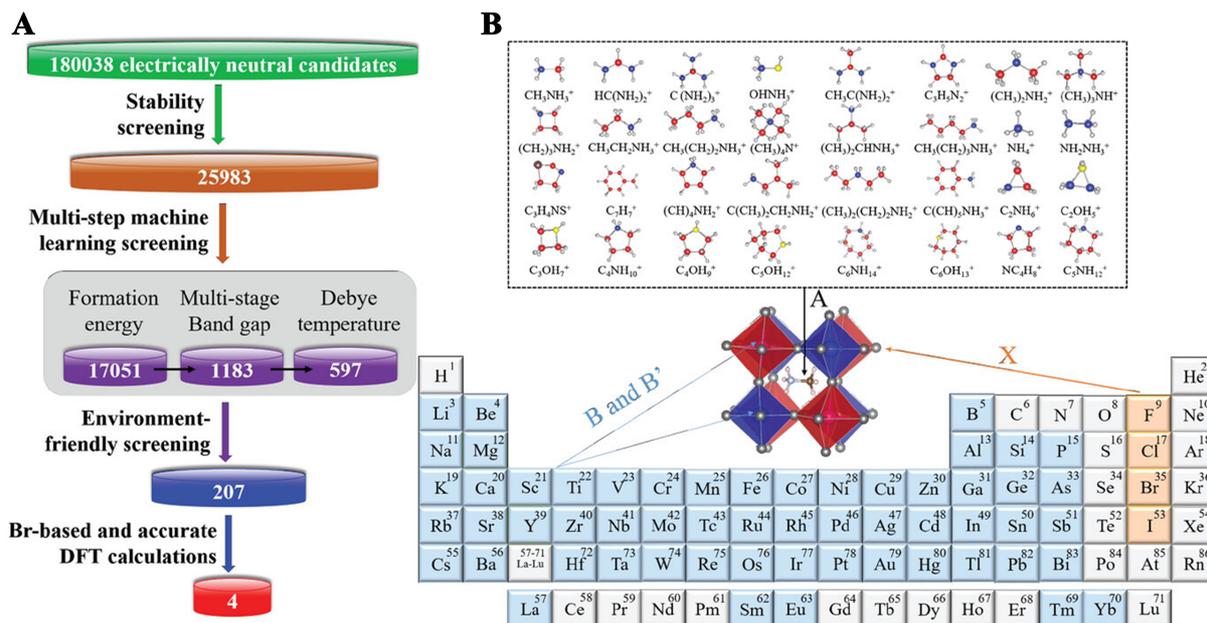


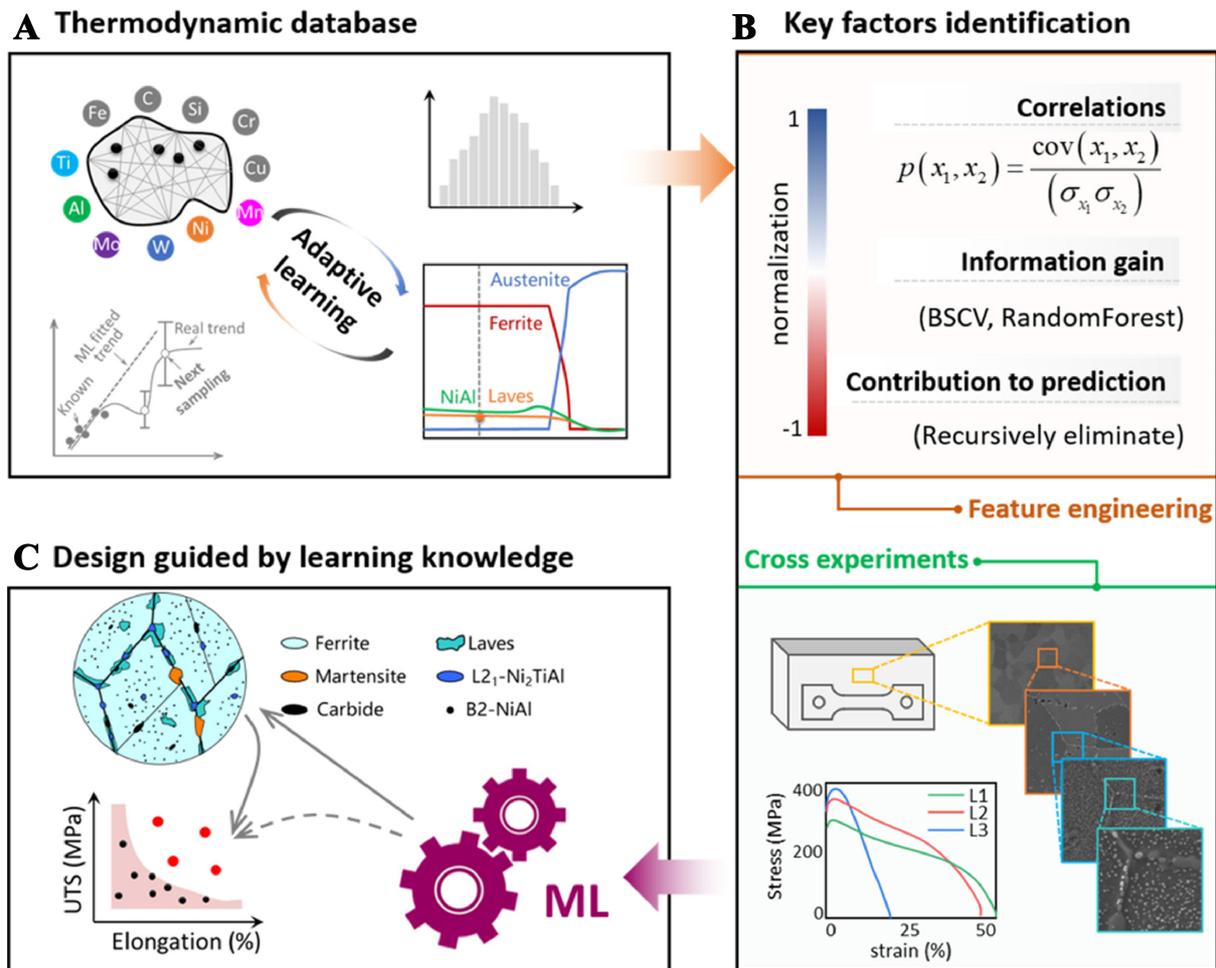
Figure 7. (A) The process of discovering novel HOIDPs according to the combination of machine learning and DFT calculation for photovoltaic application; (B) The composition and structure of perovskites in the prediction set [72]. HOIDPs: Hybrid organic-inorganic double perovskites; DFT: density functional theory.

candidates with structurally feasible compositions. Next, samples with formation energies greater than  $-0.2$  eV were excluded, leading to 17,051 structurally stable candidates. By applying PBE and HSE bandgap ranges of 0.6–2.2 and 1.1–3.0 eV, respectively, they selected 1,183 candidates suitable for light absorption and matching solar cell power conversion efficiency (PCE). Furthermore, through screening for Debye temperatures above 500 K, they identified 597 candidates with high efficiency and stability. Lastly, by

considering environmental friendliness, they obtained 207 lead-free candidates. Environmentally friendly candidates were ultimately selected based on Br and further verified using DFT. Finally, four candidates with excellent stability, high Debye temperatures, and suitable bandgaps were screened as promising materials for solar cells. Additionally, through DFT calculations, they confirmed three lead-free candidates of  $(\text{CH}_3\text{NH}_3)_2\text{AgGaBr}_6$ ,  $(\text{CH}_3\text{NH}_3)_2\text{AgInBr}_6$ , and  $(\text{C}_2\text{NH}_6)_2\text{AgInBr}_6$ . These research findings provide valuable guidance for the discovery of high-performance solar cell materials.

Gao *et al.* proposed a machine learning-driven alloy design strategy that combines feature analysis and layer-by-layer multi-objective optimization to achieve a breakthrough in the properties of a novel high-performance lightweight refractory high-entropy alloy (LW-RHEA)<sup>[73]</sup>. According to domain knowledge, the LW-RHEA with a bcc\_A2 single phase holds the potential for high hardness and high corrosion resistance. To this end, the authors amassed 92 phase structure datasets of arc-melted Al-Nb-Ti-V-Zr-Cr-Mo-Hf LW-RHEAs from the literature and utilized the alloy composition information as a feature to construct a phase structure classification model in conjunction with a radial basis function kernel support vector machine classification. The average accuracies of the training and test sets were 98.63% and 94.74%, respectively. The hardness dataset of the bcc\_A2 single-phase Al-Nb-Ti-V-Zr-Cr-Mo-Hf alloy was collected and combined with multiple linear regression to develop a hardness prediction model, with the  $R^2$  values of the training and test sets reaching as high as 0.95 and 0.92, respectively. The SHAP analysis of the features indicated that the content of the Cr element has a significant influence on the corrosion resistance of the alloy, and a Cr content greater than 12 at.% is regarded as the criterion for high corrosion resistance alloys. Three target alloys, namely  $\text{Al}_{20}\text{Nb}_{28}\text{Ti}_{20}\text{V}_4\text{Cr}_{20}\text{Mo}_8$ ,  $\text{Al}_{14}\text{Nb}_{22}\text{Ti}_{30}\text{V}_2\text{Cr}_{20}\text{Mo}_{12}$ , and  $\text{Al}_{18}\text{Nb}_{22}\text{Ti}_{34}\text{V}_4\text{Cr}_{20}\text{Mo}_{12}$ , were designed from 949,307 virtual alloy samples, with the criteria of phase structure, hardness, theoretically calculated density, and melting point falling within the high, medium, and low ranges, corresponding to high hardness, excellent comprehensive performance, and excellent corrosion resistance alloys, respectively. The experimental results demonstrated that the phase structures of the three designed alloys are single-phase bcc\_A2, and their hardness and corrosion resistance measurements are highly consistent with the predicted results.

Hu *et al.* proposed an active learning framework consisting of a three-step learning strategy in Figure 8 for designing high-strength and ductile 15Cr ferritic steels<sup>[74]</sup>. In this study, the content of the Laves phase and ferrite phase was considered as key factors influencing the strength and ductility of ferritic steels. The first step involved constructing a high-quality thermodynamic database. The research team generated 40 samples with varying compositions and step sizes of Mn, Ni, Al, Ti, Mo, and W. Using thermodynamic software, they calculated the content of Laves phase and ferrite phase for these samples. Among the 12 modeling algorithms applied to these 40 samples, ridge regression and GPR demonstrated higher predictive accuracy. To further evaluate the performance of these two models within the active learning framework, 15 samples were randomly selected from the initial 40 samples, leaving the rest as an unexplored set. They conducted 20 iterations of active learning and obtained the databases for the content of the Laves phase and ferrite phase, comprising 1,260 and 1,120 samples, respectively. The second step involved identifying key factors. Through PCC analysis, random forest importance analysis, and exhaustive elimination analysis of the database, they explored the effects of six alloying elements - Mn, Ni, Al, Ti, Mo, and W on the content of Laves phase and ferrite phase. Further, using orthogonal experimental design, they generated 16 samples and measured their UTS and elongation. The third step involved constructing a predictive model for tensile properties. After the model construction for UTS and elongation, the model analysis revealed that the addition of Mo improved both strength and ductility, and the introduction of a small amount of Ti at low W content produced similar effects. However, beyond 1 wt% Ti content, the strength continued to increase while the ductility decreased. The influence of W on strength was weaker and more complex, with its content



**Figure 8.** Three-step learning strategy for designing 15Cr ferritic steels<sup>[74]</sup>. (A) The first step: construct the thermodynamic composition-microstructure database; (B) The second step: identify the key factors influencing the Laves phase; (C) Domain knowledge extracted from experimental data guides the reasonable design of new steels.

showing an inverse relationship with ductility. Based on the domain knowledge derived from the machine learning model, the authors designed five samples with high strength and ductility, and these were experimentally verified. The experimental results demonstrated that the best sample had an UTS of 405 MPa and an elongation of 50.99%, confirming the success of the 3-step design strategy in improving strength and ductility. The highlight of this work lies in the construction of a high-quality database using the active learning framework and the consideration of predictive accuracy within the framework during algorithm evaluation. While the study did not involve Pareto front, valuable domain knowledge for materials design was gained through detailed analysis of the model and data, successfully achieving the objective of extracting knowledge from machine learning.

Wei *et al.* employed three machine learning methods of regularized linear regression, random forest, and ANN, to conduct multi-objective machine learning research on the fatigue strength, tensile strength, fracture strength, and hardness of steel materials<sup>[75]</sup>. They used a set of feature variables to describe these four target functions and ranked the importance of the feature variables using three methods: Multi-task-Lasso, feature importance, and complex correlation coefficient. Additionally, they utilized the “area method” to balance the model complexity with prediction accuracy. Through linear regression, they derived

analytical expressions based on chemical composition, preparation process parameters, and inclusion parameters, which exhibited good predictive capabilities for the four mechanical properties. On feature subsets containing 16, 8, and 14 features, the cross-validation R-values for linear regression reached maximum values of 0.9627, 0.9731, and 0.9777, respectively.

Small molecule hole transport materials (SM-HTMs) play a crucial role in the development of perovskite solar cells (PSCs) due to their simple structure and excellent optoelectronic properties. Zhou *et al.* constructed an interpretable model for designing materials with high hole mobility ( $\mu$ ) and absolute hardness ( $\eta$ )<sup>[76]</sup>. Initially, they collected 43  $\mu$  samples and 47  $\eta$  samples from publications and optimized the geometric shapes using the B3LYP method. The Dragon software generated 5,270 initial features. After feature selection and model comparison,  $\mu$  and  $\eta$  models were constructed using GBR to achieve optimal performance. The R values of LOOCV for  $\mu$  and  $\eta$  models were 0.86 and 0.84, respectively. The SHAP analysis of important features revealed that at a topological distance of 9, the presence of fewer C–O and N–O bonds was more likely to result in higher  $\mu$ ; furthermore, molecules with higher cyclization and fewer polar fragments possibly had higher  $\eta$ . Based on these strategies, they designed 785 virtual samples and conducted further analysis by combining model predictions. Finally, through virtual screening and pattern recognition, they discovered 35 candidates with high  $\mu$  and  $\eta$  ( $\mu > 10^{-1.54} \text{ cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$ ,  $\eta > 2.71 \text{ eV}$ ). This study provides an effective approach to accelerate the design of outstanding SM-HTMs.

Liu *et al.* proposed a generalized active learning algorithm for three-objective performance of materials named equiprobability distribution with maximum confidence interval (ED-MCI), which can be extended to co-optimize the design of performance of more interrelated objectives<sup>[77]</sup>. The algorithm constructs a probability density function for a trivariate Gaussian distribution with the expression:

$$N(y|u, \sigma) = \frac{1}{2\pi^{\frac{3}{2}}\sigma_1\sigma_2\sigma_3} \exp\left(-\frac{1}{2}\left(\frac{(y_1 - u_1)^2}{\sigma_1^2} + \frac{(y_2 - u_2)^2}{\sigma_2^2} + \frac{(y_3 - u_3)^2}{\sigma_3^2}\right)\right)$$

where  $y_i$  is the  $i$ th optimization objective;  $u_i$  is the model predicted mean of the  $i$ th optimization objective; and  $\sigma_i$  is the model predicted standard deviation of the  $i$ th optimization objective. The 3D Gaussian probability density function has an ellipsoid of equal probability distribution with a maximum confidence interval (95%). In order to make the values of the three objective attributes along the desired specific optimization direction, a weight function is used to recommend the desired candidates from the virtual sample. The function is responsible for controlling the specific optimization direction in recommending the best candidate, which is derived from an equal probability distribution with a maximum confidence interval in the multivariate Gaussian probability density function, as given by

$$W(y_1, y_2, y_3) = \frac{k_1 y_1 * k_2 y_2 * k_3 y_3}{k_1 y_1 * k_2 y_2 + k_2 y_2 * k_3 y_3 + k_1 y_1 * k_3 y_3}$$

where  $k_i$  the weight coefficient of the  $i$ th optimization objective. By calculating the weight function value of each point on the surface of the ellipsoid one by one, the candidate sample with the highest weight function value is recommended to be used for synthesis and characterization. ED-MCI was successfully applied to the design of high-temperature alloys with high temperature carrying capacity and long creep life. The machine learning model was trained with multiple small data sample sets, and through the feedback iteration of the model recommendation and experimental validation, the high-temperature alloys with high

$\gamma'$ -phase were co-optimized using ED-MCI. The CoNiAlCr-based high-temperature alloys with high  $\gamma'$ -phase volume fraction, small size and high cubicity are prepared by synergistic optimization of  $\gamma'$ -phase volume fraction, size and shape using ED-MCI.

Wang *et al.* used an interpretable multi-objective optimization approach intended to optimize the ductility as well as the thermoelectric properties of thermoelectric materials<sup>[78]</sup>. They collected 74 ABX<sub>3</sub>-type thermoelectric perovskites from previous research results, calculated Pugh's ratio with the help of elastic property automated calculation code (EPAC), and applied DFT method to calculate dimensionless quality factor ( $zT$ ) of the materials. The atomic parameters are used as feature inputs, and the PCC analysis and recursive feature elimination strategy are combined with various modeling algorithms to carry out feature screening. The results showed that the random forest regression model with 11 features presented the highest accuracy in predicting the Prandtl ratio, while the XGBoost algorithm based on eight features had the best performance in predicting the  $zT$ , with R<sup>2</sup> values of 0.98 and 0.80 for LOOCV, respectively. Subsequently, two interpretable machine learning techniques, SHAP and SISSO, are used to analyze the key features and target attributes in depth. Both SHAP and SISSO analyses show that  $EN_{(ab) A/B}$  (the ratio of the absolute electronegativity of the elements in the A and B positions) is positively correlated with the  $zT$ , and negatively correlated with the Prandtl ratio. Further investigation of the relationship between  $EN_{(ab) A/B}$  and the conduction band bottom condensation ( $N_c$ ) reveals that compounds with higher  $EN_{(ab) A/B}$  values tend to have higher energy band condensation, which further contributes to the higher  $zT$  values. In addition, higher  $EN_{(ab) A/B}$  corresponds to weaker B-X bonding, which not only reduces the shear modulus  $G$  and the Prandtl ratio of the material, but also enhances the ductility of the material. By integrating the interpretable machine learning technique with the multi-objective optimization strategy, this study provides valuable ideas and methods to accelerate the exploitation of thermoelectric materials with high performance and good ductility.

The details of various research advancements are presented in [Supplementary Table 3](#). It can be observed that researchers tend to collect data and construct models separately for each optimization objective, regardless of whether Pareto front is considered. In cases where Pareto front is considered, a multi-objective optimization strategy based on Pareto front is employed. This strategy utilizes models, virtual samples, and multi-objective optimization algorithms to determine the Pareto front. Additionally, the algorithms used for calculating the Pareto front predominantly involve multi-objective genetic algorithms. On the other hand, in scenarios where Pareto front is not considered, optimization objectives are tailored according to the materials application requirements. Models are then applied through a stepwise virtual sample screening process.

It can be noticed that the virtual sample generation technique occupies a very important place in multi-objective optimization. How to maintain the relationship between material structure and properties in the process of generating virtual samples and avoid generating physically unrealistic virtual samples has been a hot research topic in materials machine learning. Li *et al.* proposed the cardiGAN model, which combines generative adversarial networks (GANs) and discriminative NNs, to generate new alloy compositions approximating the distribution of the training data and predicting their phase structure using existing MPEAs data training<sup>[79]</sup>. Iyer *et al.* coped with the problem of small dataset of corrosion inhibitors using the virtual sample generation approach of the conditional tabular GAN (CTGAN) algorithm and experimentally verified the credibility and generality of the proposed ANN + VSG model<sup>[80]</sup>. Two new corrosion inhibitor compounds based on 2-alkylbenzimidazole scaffolds were synthesized and their corrosion inhibition efficiencies were experimentally obtained, and their experimental values were in high agreement with the model predictions. CTGAN's embedding of the synthesized data in the training samples

enhanced the model's ability to identify feature-target relationships, thereby stabilizing and improving the correlation between the chemical quantum descriptors and inhibition efficiencies.

## DISCUSSION

In the multi-objective optimization strategy for obtaining the Pareto front, the Pareto optimal solutions can be obtained by a multi-objective optimization algorithm or by generating a large number of virtual samples and performing non-dominated sorting on their target properties prediction values. This part aims to provide a detailed discussion of different Pareto optimal solutions calculation strategies based on the previous work in our lab.

As described in the work by Ma<sup>[62]</sup>, we have developed predictive models (LightGBM-H, SVR-D) for the  $H$  and  $D$  of high-entropy alloys. We merged their best feature subsets and rebuilt the prediction models (LightGBM-H2, SVR-D2) to ensure that both target variables share the same features, thereby establishing a connection between them. These prediction functions were then used as the fitness functions for NSGA-II, with individuals (i.e., candidate solutions to the optimization problem) represented by sequences of feature values. To back-calculate the compositions of the alloys, we generated 100,000 virtual samples of chemical formulas and computed their feature values. By calculating the weighted Euclidean distance between the feature value sequences of the virtual samples and those on the Pareto front (where weights are proportional to the SHAP importance score of each feature in the unmerged feature models), we selected the three closest virtual samples as approximations of each Pareto optimal solution represented by each feature value sequence. This Pareto optimal solution calculation strategy is referred to as Strategy 1.

In multi-objective optimization algorithms, it is essential for the multiple objectives to exhibit some form of interrelation. If the objectives are completely unrelated, the problem essentially reduces to multiple single-objective optimization tasks, making multi-objective optimization algorithms unsuitable. Such interrelations can manifest as consistent trends between objectives, conflicting relationships, or even more complex interactions. These relationships are established through shared variable sequences within the multi-objective optimization algorithm, where updates to the variable sequence values concurrently influence the predicted values of both objective variables. Strategy 1 selects the modeling features of the target properties as the genes for NSGA-II. However, since the best feature subsets for  $H$  and  $D$  are not consistent, merging these subsets is necessary. The continued high accuracy achieved using the merged features indicates that these features impact both target properties, despite not being included in the best subsets with fewer features. However, there are limitations: after merging, the model's predictive accuracy might decrease, and with an increase in the number of features, there is a higher risk of overfitting, especially when data is sparse. Additionally, virtual screening is needed to approximate the Pareto front, and there may be discrepancies between the objective values of these approximations and the actual Pareto front.

To improve Strategy 1, Strategy 2 is proposed. Unlike Strategy 1, where the feature variable sequences are used as the chromosomes in NSGA-II, Strategy 2 uses the sequence of element content as the chromosomes. Each individual (candidate solution) is a 23-dimensional vector corresponding to the chemical composition of an alloy. The chemical composition is then used to compute the corresponding features, which are input into the predictive models (LightGBM-H, SVR-D) to obtain the predicted target properties. Since the features are derived from the alloy's chemical composition, they can be considered intermediate variables that link the predictions of multiple target properties to the chemical formula. Both Strategy 1 and Strategy 2 employed NSGA-II. To explore its advantages in solving multi-objective optimization tasks, we introduced Strategy 3, which does not utilize NSGA-II. Strategy 3 involves directly generating 100,000 virtual samples of chemical formulas and their features, followed by using the models (LightGBM-H, SVR-

D) for predictions. The non-dominated sorting is then used to calculate the Pareto front of the virtual samples.

To ensure a fair comparison of the three strategies under consistent conditions, we adjusted certain parameters of NSGA-II and the conditions for generating virtual samples, relative to the original study. For both Strategy 1 and Strategy 2, the NSGA-II population size was set to 200, with a maximum number of iterations of 500, which resulted in a total of 100,000 individuals. The number of virtual samples generated by both Strategy 1 and Strategy 3 is 100,000, matching the number of individuals generated by NSGA-II. The constraints for all three strategies, including the probability of element selection, atomic percentage range, and the number of constituent elements, are kept consistent. Setting constraints on element selection probability and atomic percentages ensures that the individuals in NSGA-II and the virtual samples closely resemble the samples in the original datasets, thereby reducing the risk of model predictions for unknown samples. The selection probability of each element is determined by its average frequency in both datasets. In the virtual sample generation for Strategies 1 and 3, the element composition of the virtual samples is randomly generated using a roulette wheel selection algorithm. For Strategy 2, the range of element content in NSGA-II was expanded based on the corresponding range in the datasets. If the generated content falls within the range consistent with the original dataset, the element is considered to be present with the content equal to the generated value. If it is outside the content range of the original dataset, the element is considered absent in the alloy. The ratio of the length of the presence interval to the total interval length is proportional to the probability of selecting the element. Elements with a higher selection probability are more likely to be chosen. Conversely, elements with a lower selection probability still have a chance of being selected, thus preventing the direct elimination of low-frequency elements and enhancing the diversity of the virtual samples (population). For the atomic percentage constraints, all three strategies generate initial content within the atomic percentage range corresponding to each element in the dataset. The content is then divided by the total content of all elements to obtain the atomic percentage. To ensure that the atomic percentages remain within the range of the original dataset, both the virtual sample generation method and NSGA-II constrain the total content of all elements to lie between 0.9 and 1.1.

The Pareto fronts of the three strategies are shown in [Figure 9](#). Given the objective of selecting alloys with higher  $H$  and  $D$ , Strategy 2 is identified as the optimal strategy. The method of strategy 1 is to obtain the Pareto optimal solution represented by a sequence of eigenvalues and then back-calculate the combinations based on the distances, which may lead to irrational combinations of features, making it difficult to find well-matching compositions. Strategy 3 overly relies on virtual sample generation techniques; insufficient quantity or rationality of virtual samples could significantly impact the Pareto front calculation. In contrast, Strategy 2 avoids the issues of feature merging and back-calculation in Strategy 1. Under the condition that the total number of individuals and virtual samples is the same, Strategies 1 and 2 achieved superior Pareto fronts compared to Strategy 3, highlighting the advantages of multi-objective evolutionary algorithms over virtual screening. By leveraging NSGA-II, the evolution of the population is not entirely random but inherits the advantages of previous generations, thereby forming a more effective evolutionary direction and progressively achieving individuals with higher fitness, rather than relying solely on an accumulation of virtual samples to obtain a better Pareto front in Strategy 3. Notably, if there are no constraints on the selection probability and content of elements, a better Pareto front might be obtained, but with greater predictive uncertainty. Additionally, we have only considered the probability of individual elements being selected, without addressing the joint probabilities of multiple element combinations. Further exploration of virtual sample generation strategies that more closely resemble and realistically represent the dataset is warranted.

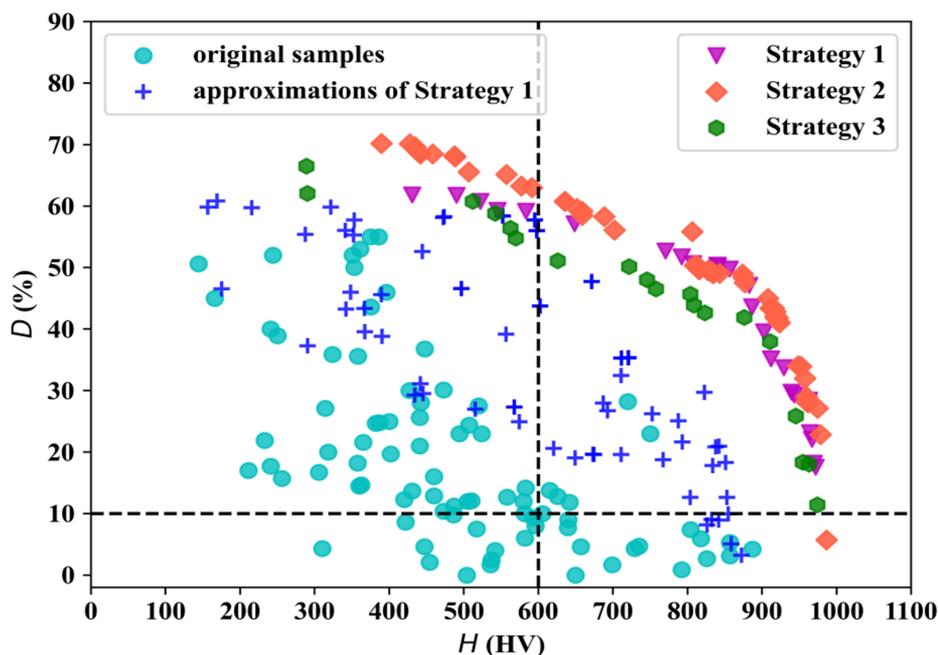


Figure 9. The comparison of the three strategies.

## CONCLUSION AND OUTLOOKS

This review aims to explore the cutting-edge research achievements in the multi-objective optimization of machine learning-assisted materials design and discovery. We present the materials machine learning process based on multi-objective optimization and discuss the application of Pareto front and its computation methods in scenarios involving trade-offs among objectives. Additionally, we provide an overview of common multi-objective optimization strategies and summarize cases of multi-objective optimization research in the field of materials machine learning. In the future, we propose the development directions from the following perspectives:

(1) Cross-scale and multi-physics field coupling optimization: The properties of materials are often affected by multiple levels of structures and factors from an atomic to a macroscopic scale, so it is crucial to establish a cross-scale multi-objective optimization model. By combining different scales of simulation methods such as quantum mechanics, molecular dynamics, and finite element analysis, we explore the features at different scales and combine them with machine learning modeling to achieve comprehensive optimization of materials from microstructure to macroscopic properties. Materials in practical applications usually involve the interaction of multiple physical fields, such as force, heat, electricity and magnetism. Therefore, it is necessary to develop a multi-objective optimization method with multi-physical field coupling to realize the comprehensive performance optimization of materials under complex working conditions.

(2) Multi-objective reverse design: Traditional material design usually starts from the composition and process, and predicts the performance of the material positively, while material reverse design starts from the desired material performance, and deduces the composition, microstructure and preparation process of the material negatively. Currently, machine learning-based inverse design methods, such as genetic algorithms, pattern recognition inverse projection, and progressive active search, have relatively mature applications in materials design of polymer, alloy, and perovskites. However, most of these applications focus on single-objective optimization, and in the future, the inverse design algorithms can be further

enhanced and substantially applied to the inverse design of materials with multi-objective properties.

(3) Multi-source data fusion: In material machine learning, the scarcity of data has always been a major constraint on the accuracy of the model, although there are methods such as migration learning, active learning, high-throughput computation and experimentation to solve the problem of small samples, it has always been a challenge for scientifically fusing the data from different sources to improve the data volume for modeling. In the future, data volume can be improved by introducing the concept of fidelity to fuse data from different sources. Multi-fidelity modeling is a kind of model that combines different fidelity data to construct a model, aiming at obtaining a model whose accuracy meets the design requirements with as little high-cost high-fidelity data as possible. When the amount of experimental data is limited, the fidelity model can be used to fuse experimental data with computational data to improve the accuracy and application value of the model and speculation.

## DECLARATIONS

### Authors' contributions

Proposed and conceptualized the theme and framework of this review; collected and analyzed the literature; and wrote the manuscript: Xu, P.

Coded and wrote the "Discussion" section: Ma, Y.; Lu, W.

Supervised and revised the manuscript: Li, M.; Zhao, W.; Dai, Z.

### Availability of data and materials

All the data of the cases could be obtained from the corresponding references. The code and data files of the three strategies in the discussion are available at <https://gitee.com/yingyingdaydayup/comparison-of-3-multi-objective-strategies.git>.

### Financial support and sponsorship

None.

### 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) 2025.

## REFERENCES

1. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP. J. Adv. Signal. Process.* **2016**, *2016*, 67. DOI
2. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547-55. DOI PubMed
3. Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. *npj. Comput. Mater.* **2023**, *9*, 1000. DOI
4. Kumari, J.; Kumar, E.; Kumar, D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. *Arch. Comput. Methods. Eng.* **2023**, *30*, 3673-701. DOI PubMed PMC
5. An, Q.; Rahman, S.; Zhou, J.; Kang, J. J. A comprehensive review on machine learning in healthcare industry: classification,

- restrictions, opportunities and challenges. *Sensors* **2023**, *23*, 4178. DOI PubMed PMC
6. Nazareth, N.; Ramana Reddy, Y. V. Financial applications of machine learning: a literature review. *Expert. Syst. Appl.* **2023**, *219*, 119640. DOI
  7. Ghoddusi, H.; Creamer, G. G.; Rafizadeh, N. Machine learning in energy economics and finance: a review. *Energy. Econ.* **2019**, *81*, 709-27. DOI
  8. Kim, D. H.; Kim, T. J. Y.; Wang, X.; et al. Smart machining process using machine learning: a review and perspective on machining industry. *Int. J. Precis. Eng. Manuf. Green. Technol.* **2018**, *5*, 555-68. DOI
  9. Mccooy, J.; Auret, L. Machine learning applications in minerals processing: a review. *Min. Eng.* **2019**, *132*, 95-109. DOI
  10. Kadulkar, S.; Sherman, Z. M.; Ganesan, V.; Truskett, T. M. Machine learning-assisted design of material properties. *Annu. Rev. Chem. Biomol. Eng.* **2022**, *13*, 235-54. DOI PubMed
  11. 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
  12. Si, Z.; Zhou, D.; Yang, J.; Lin, X. Review: 2D material property characterizations by machine-learning-assisted microscopies. *Appl. Phys. A.* **2023**, *129*, 6543. DOI
  13. Wei, J.; Chu, X.; Sun, X.; et al. Machine learning in materials science. *InfoMat* **2019**, *1*, 338-58. DOI
  14. Liu, Y.; Niu, C.; Wang, Z.; et al. Machine learning in materials genome initiative: a review. *J. Mater. Sci. Technol.* **2020**, *57*, 113-22. DOI
  15. Wang, X.; Xiao, R.; Li, H.; Chen, L. Discovery and design of lithium battery materials via high-throughput modeling. *Chinese. Phys. B.* **2018**, *27*, 128801. DOI
  16. Hu, W.; Zhang, L. First-principles, machine learning and symbolic regression modelling for organic molecule adsorption on two-dimensional CaO surface. *J. Mol. Graph. Model.* **2023**, *124*, 108530. DOI
  17. Mai, J.; Lu, T.; Xu, P.; Lian, Z.; Li, M.; Lu, W. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. *Dyes. Pigments.* **2022**, *206*, 110647. DOI
  18. Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. *npj. Comput. Mater.* **2021**, *7*, 495. DOI
  19. Wang, J.; Xu, P.; Ji, X.; Li, M.; Lu, W. Feature selection in machine learning for perovskite materials design and discovery. *Materials* **2023**, *16*, 3134. DOI PubMed PMC
  20. Bhat, N.; Barnard, A. S.; Birbilis, N. Unsupervised machine learning discovers classes in aluminium alloys. *R. Soc. Open. Sci.* **2023**, *10*, 220360. DOI PubMed PMC
  21. 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
  22. Sha, W.; Li, Y.; Tang, S.; et al. Machine learning in polymer informatics. *InfoMat* **2021**, *3*, 353-61. DOI
  23. Xu, P.; Chen, H.; Li, M.; Lu, W. New opportunity: machine learning for polymer materials design and discovery. *Adv. Theory. Simul.* **2022**, *5*, 2100565. DOI
  24. Startt, J.; Mccarthy, M. J.; Wood, M. A.; Donegan, S.; Dingreville, R. Bayesian blacksmithing: discovering thermomechanical properties and deformation mechanisms in high-entropy refractory alloys. *npj. Comput. Mater.* **2024**, *10*, 1353. DOI
  25. Zhang, Q.; Hu, Y.; Li, S.; et al. Recent advances in supported acid/base ionic liquids as catalysts for biodiesel production. *Front. Chem.* **2022**, *10*, 999607. DOI PubMed PMC
  26. Zheng, M.; Wang, Y.; Teng, H. A new "intersection" method for multi-objective optimization in material selection. *Teh. glas.* **2021**, *15*, 562-8. DOI
  27. Al Ani, Z.; Gujarathi, A. M.; Al-Muhtaseb, A. H. A state of art review on applications of multi-objective evolutionary algorithms in chemicals production reactors. *Artif. Intell. Rev.* **2023**, *56*, 2435-96. DOI
  28. Takagi, T.; Takadama, K.; Sato, H. Directional pareto front and its estimation to encourage multi-objective decision-making. *IEEE. Access.* **2023**, *11*, 20619-34. DOI
  29. Lee, J.; Lee, S.; Ahn, J.; Choi, H. Pareto front generation with knee-point based pruning for mixed discrete multi-objective optimization. *Struct. Multidisc. Optim.* **2018**, *58*, 823-30. DOI
  30. Shi, L.; Chang, D.; Ji, X.; Lu, W. Using data mining to search for perovskite materials with higher specific surface area. *J. Chem. Inf. Model.* **2018**, *58*, 2420-7. DOI
  31. Chen, H.; Shang, Z.; Lu, W.; Li, M.; Tan, F. A property-driven stepwise design strategy for multiple low-melting alloys via machine learning. *Adv. Eng. Mater.* **2021**, *23*, 2100612. DOI
  32. Zhang, S.; Lu, T.; Xu, P.; Tao, Q.; Li, M.; Lu, W. Predicting the formability of hybrid organic-inorganic perovskites via an interpretable machine learning strategy. *J. Phys. Chem. Lett.* **2021**, *12*, 7423-30. DOI
  33. Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **2019**, *5*, eaav0693. DOI PubMed PMC
  34. Wang, T.; Hu, J.; Ouyang, R.; et al. Nature of metal-support interaction for metal catalysts on oxide supports. *Science* **2024**, *386*, 915-20. DOI
  35. 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. DOI
  36. Li, Y.; Li, T.; Liu, H. Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* **2017**, *53*, 551-77. DOI

37. Xie, J.; Sage, M.; Zhao, Y. F. Feature selection and feature learning in machine learning applications for gas turbines: a review. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105591. DOI
38. Borchers, A.; Pieler, T. Programming pluripotent precursor cells derived from *Xenopus* embryos to generate specific tissues and organs. *Genes* **2010**, *1*, 413-26. DOI PubMed PMC
39. Zheng, W.; Chen, S.; Fu, Z.; Zhu, F.; Yan, H.; Yang, J. Feature selection boosted by unselected features. *IEEE. Trans. Neural. Netw. Learn. Syst.* **2022**, *33*, 4562-74. DOI
40. 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
41. Wang, J.; Xu, P.; Ji, X.; Li, M.; Lu, W. MIC-SHAP: an ensemble feature selection method for materials machine learning. *Mater. Today. Commun.* **2023**, *37*, 106910. DOI
42. Ordillo, V. Z.; Shimizu, K.; Putungan, D. B.; et al. Two-stage feature selection for machine learning-aided DFT-based surface reactivity study on single-atom alloys. *Modelling. Simul. Mater. Sci. Eng.* **2024**, *32*, 065003. DOI
43. Lu, K.; Chang, D.; Ji, X.; Li, M.; Lu, W. Machine learning aided discovery of the layered double hydroxides with the largest basal spacing for super-capacitors. *International. J. Electrochem. Sci.* **2021**, *16*, 211146. DOI
44. Zhao, Y.; Zhang, J.; Xu, Z.; et al. Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. *Nat. Commun.* **2021**, *12*, 2191. DOI PubMed PMC
45. Wong, T. T.; Yeh, P. Y. Reliable accuracy estimates from *k*-fold cross validation. *IEEE. Trans. Knowl. Data. Eng.* **2020**, *32*, 1586-94. DOI
46. Zhang, J.; Wang, S. A fast leave-one-out cross-validation for SVM-like family. *Neural. Comput. Appl.* **2016**, *27*, 1717-30. DOI
47. Meiyazhagan, J.; Sudharsan, S.; Venkatesan, A.; Senthilvelan, M. Prediction of occurrence of extreme events using machine learning. *Eur. Phys. J. Plus.* **2022**, *137*, 2249. DOI
48. Novello, P.; Poëtto, G.; Lugato, D.; Congedo, P. M. Goal-oriented sensitivity analysis of hyperparameters in deep learning. *J. Sci. Comput.* **2023**, *94*, 2083. DOI
49. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput. Environ. Urban. Syst.* **2022**, *96*, 101845. DOI
50. Szepannaek, G.; Lübke, K. How much do we see? On the explainability of partial dependence plots for credit risk scoring. *Argum. Oeconom.* **2023**, *2023*, 137-50. DOI
51. Feng, W.; Gong, D.; Yu, Z. Multi-objective evolutionary optimization based on online perceiving Pareto front characteristics. *Inform. Sci.* **2021**, *581*, 912-31. DOI
52. Khorram, E.; Khaledian, K.; Khaledyan, M. A numerical method for constructing the Pareto front of multi-objective optimization problems. *J. Comput. Appl. Math.* **2014**, *261*, 158-71. DOI
53. Bejarano, L. A.; Espitia, H. E.; Montenegro, C. E. Clustering analysis for the Pareto optimal front in multi-objective optimization. *Computation* **2022**, *10*, 37. DOI
54. Gomes, G. F.; de Almeida, F. A.; da Silva Lopes Alexandrino, P.; da Cunha, S. S.; de Sousa, B. S.; Ancelotti, A. C. A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation. *Eng. Comput.* **2019**, *35*, 519-35. DOI
55. Boş, R. I.; Grad, S.; Wanka, G. A general approach for studying duality in multiobjective optimization. *Math. Meth. Oper. Res.* **2007**, *65*, 417-44. DOI
56. Jiménez, B.; Novo, V.; Vilchez, A. A set scalarization function based on the oriented distance and relations with other set scalarizations. *Optimization* **2018**, *67*, 2091-116. DOI
57. Bazgan, C.; Ruzika, S.; Thielen, C.; Vanderpooten, D. The power of the weighted sum scalarization for approximating multiobjective optimization problems. *Theory. Comput. Syst.* **2022**, *66*, 395-415. DOI
58. Hanaoka, K. Comparison of conceptually different multi-objective Bayesian optimization methods for material design problems. *Mater. Today. Commun.* **2022**, *31*, 103440. DOI
59. Qu, B. Y.; Suganthan, P. N. Constrained multi-objective optimization algorithm with an ensemble of constraint handling methods. *Eng. Optim.* **2011**, *43*, 403-16. DOI
60. Kazemzadeh Azad, S.; Aminbakhsh, S.  $\epsilon$ -constraint guided stochastic search with successive seeding for multi-objective optimization of large-scale steel double-layer grids. *J. Build. Eng.* **2022**, *46*, 103767. DOI
61. Li, X.; Shan, G.; Zhang, J.; Shek, C. Accelerated design for magnetic high entropy alloys using data-driven multi-objective optimization. *J. Mater. Chem. C.* **2022**, *10*, 17291-302. DOI
62. Ma, Y.; Li, M.; Mu, Y.; Wang, G.; Lu, W. Accelerated design for high-entropy alloys based on machine learning and multiobjective optimization. *J. Chem. Inf. Model.* **2023**, *63*, 6029-42. DOI
63. Chen, Y.; Tian, Y.; Zhou, Y.; et al. Machine learning assisted multi-objective optimization for materials processing parameters: a case study in Mg alloy. *J. Alloys. Compd.* **2020**, *844*, 156159. DOI
64. Deng, Y.; Zhang, Y.; Gong, X.; et al. An intelligent design for Ni-based superalloy based on machine learning and multi-objective optimization. *Mater. Design.* **2022**, *221*, 110935. DOI
65. Banerjee, T.; Dey, S.; Sekhar, A. P.; Datta, S.; Das, D. Design of alumina reinforced aluminium alloy composites with improved tribo-mechanical properties: a machine learning approach. *Trans. Indian. Inst. Met.* **2020**, *73*, 3059-69. DOI
66. Wu, S.; Xu, X.; Yang, S.; Qiu, J.; Volinsky, A. A.; Pang, X. Data-driven optimization of hardness and toughness of high-entropy

- nitride coatings. *Ceram. Int.* **2023**, *49*, 21561-9. [DOI](#)
67. Zhang, Y.; Xie, S.; Guo, W.; Ding, J.; Poh, L. H.; Sha, Z. Multi-objective optimization for high-performance Fe-based metallic glasses via machine learning approach. *J. Alloys. Compd.* **2023**, *960*, 170793. [DOI](#)
  68. Gao, W.; Wang, B.; Gu, Q.; et al. Accelerated discovery of high-performance 3D printing materials using multi-objective active optimization method. *J. Mater. Sci.* **2024**, *59*, 2390-402. [DOI](#)
  69. Morales-Hernández, A.; Rojas Gonzalez, S.; Van Nieuwenhuysse, I.; et al. Bayesian multi-objective optimization of process design parameters in constrained settings with noise: an engineering design application. *Eng. Comput.* **2024**, *40*, 2497-511. [DOI](#)
  70. Fang, J.; Xie, M.; Zhang, J.; et al. Optimized design of composition and brazing process for Cu-Ag-Zn-Mn-Ni-Si-B-P alloy brazing material based on machine learning strategy to improve brazing properties. *Mater. Today. Commun.* **2024**, *39*, 109317. [DOI](#)
  71. Tao, Q.; Chang, D.; Lu, T.; et al. Multiobjective stepwise design strategy-assisted design of high-performance perovskite oxide photocatalysts. *J. Phys. Chem. C.* **2021**, *125*, 21141-50. [DOI](#)
  72. Cai, X.; Zhang, Y.; Shi, Z.; et al. Discovery of lead-free perovskites for high-performance solar cells via machine learning: ultrabroadband absorption, low radiative combination, and enhanced thermal conductivities. *Adv. Sci.* **2022**, *9*, e2103648. [DOI](#) [PubMed](#) [PMC](#)
  73. Gao, T.; Gao, J.; Yang, S.; Zhang, L. Data-driven design of novel lightweight refractory high-entropy alloys with superb hardness and corrosion resistance. *npj. Comput. Mater.* **2024**, *10*, 1457. [DOI](#)
  74. Hu, X.; Chen, Y.; Lu, J.; et al. Three-step learning strategy for designing 15Cr ferritic steels with enhanced strength and plasticity at elevated temperature. *J. Mater. Sci. Technol.* **2023**, *164*, 79-94. [DOI](#)
  75. Wei, Q.; Xiong, J.; Sun, S.; Zhang, T. Multi-objective machine learning of four mechanical properties of steels. *Sci. Sin. Tech.* **2021**, *51*, 722-36. [DOI](#)
  76. Zhou, X.; Zheng, Z.; Lu, T.; et al. Interpretable machine learning assisted multi-objective optimization design for small molecule hole transport materials. *J. Alloys. Compd.* **2023**, *966*, 171440. [DOI](#)
  77. Liu, P.; Huang, H.; Wen, C.; Lookman, T.; Su, Y. The  $\gamma/\gamma'$  microstructure in CoNiAlCr-based superalloys using triple-objective optimization. *npj. Comput. Mater.* **2023**, *9*, 1090. [DOI](#)
  78. Wang, X.; Cao, Y.; Ji, J.; Sheng, Y.; Yang, J.; Ke, X. A multi-objective, multi-interpretable machine learning demonstration verified by domain knowledge for ductile thermoelectric materials. *J. Materiomics.* **2025**, *11*, 100886. [DOI](#)
  79. Li, Z.; Nash, W.; O'Brien, S.; Qiu, Y.; Gupta, R.; Birbilis, N. cardiGAN: a generative adversarial network model for design and discovery of multi principal element alloys. *J. Mater. Sci. Technol.* **2022**, *125*, 81-96. [DOI](#)
  80. Iyer, R. S.; Iyer, N. S.; P, R. A.; Joseph, A. Harnessing machine learning and virtual sample generation for corrosion studies of 2-alkyl benzimidazole scaffold small dataset with an experimental validation. *J. Mol. Struct.* **2024**, *1306*, 137767. [DOI](#)