Research Article

Material intelligent visualization design via two-dimensional symbolic feature generation

Wei Yong[1], Hongtao Zhang[1,2,3,5,*], Zhuang Li[1], Jie He[1], Chubo Chen[1], Yaxin Gao[1], Huadong Fu[1,2,3,4,5,*], Jianxin Xie[1,2,3,4,5]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China.
[2]Beijing Key Laboratory of Materials Intelligent Technology, School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China.
[3]Beijing Laboratory of Metallic Materials and Processing for Modern Transportation, School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China.
[4]Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, Liaoning, China.
[5]Key Laboratory for Advanced Materials Processing (MOE), School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China.

[*]**Correspondence to:** Prof. Hongtao Zhang, Prof. Huadong Fu, Beijing Advanced Innovation Center for Materials Genome Engineering, School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China. E-mail: zht@ustb.edu.cn; hdfu@ustb.edu.cn

How to cite this article: Yong W, Zhang H, Li Z, He J, Chen C, Gao Y, Fu H, Xie J. Material intelligent visualization design via two-dimensional symbolic feature generation. *J Mater Inf* 2026;6:[Accept]. http://dx.doi.org/10.20517/jmi.2026.02

Received: 27 January 2026 | **Revised:** 11 March 2026 | **Accepted:** 26 March 2026

Abstract

Machine learning for complex materials problems suffers from high-dimensionality data, while traditional "black-box" dimensionality reduction techniques typically lack the capability to balance predictive accuracy with visualization and interpretability. This work presents a novel method named Two-Dimensional Symbolic Feature Generation, based on symbolic regression and genetic algorithms. This approach facilitates machine learning by providing quantifiable interpretability and enabling visualization-driven materials design. Evaluated across diverse classification and regression tasks in materials science, the proposed method demonstrates notable success in three critical aspects: First, it significantly improved the predictive accuracy. Specifically, classification accuracy for ferroelectric perovskites and high-entropy alloy phases improved from 85.6% and 84.5% to 94.2% and 88.4%, respectively. Correspondingly, prediction errors for shape memory alloys and copper alloys were reduced from 2.9 K, 5.9% and 9.7% to 1.1 K, 3.7% and 6.2%. Second, the method ensures interpretability by constructing explicit mathematical expressions that transform the original high-dimensional features into two new symbolic ones, avoiding opaque spatial transformations. Third, it enables model visualization through two-dimensional contour maps that relate the constructed features to the target property, thereby offering intuitive insight into feature-property relationships. Leveraging these "design roadmaps," a refractory high-entropy alloy with a single-phase solid solution and a precipitation-strengthened copper alloy with an optimized property trade-off were successfully designed. The two-dimensional symbolic feature generation framework thus addresses key limitations in accuracy, interpretability, and visualization within materials informatics, establishing a new paradigm for transparent and visual materials design.

Keywords: Machine learning, feature construction, interpretability, materials visualization design

INTRODUCTION

In recent years, progress in artificial intelligence has empowered machine learning (ML) to achieve notable successes in materials development[1-6]. ML techniques have been effectively employed in composition design, process optimization and property prediction across a variety of material systems, such as copper alloys[7-11], high-entropy

alloys[12-16], superalloys[17-21], steels[22-25] and ferroelectric materials[26-30], thereby pioneering a new paradigm for accelerated materials research and development.

However, the pursuit of higher predictive accuracy in materials informatics is paradoxically leading to increasingly complex and large-scale models. Representative approaches such as large language models[31-34] and deep reinforcement learning[6,21,35-37] excel in composition design and property prediction. However, their massive parameters and multi-layered nonlinear transformations render them inherently uninterpretable 'black boxes'. Consequently, these models operate via end-to-end input-output mappings without offering navigable guidance for materials design, effectively forcing blind exploration in a high-dimensional space and thereby limiting both interpretability and the extraction of reusable design routes.

To address the interpretability issues of complex models, mainstream analytical methods such as SHapley Additive exPlanations (SHAP)[38] and Local Interpretable Model-agnostic Explanations (LIME)[39] are commonly employed to evaluate feature importance. Alibagheri *et al.*[40] developed a predictive model connecting electronic structure features to formation energy and subsequently applied SHAP analysis to rank the feature contributions, revealing a negative correlation between average ionic charge and formation energy. Similarly, Xu *et al.*[41] employed SHAP to evaluate the influence of elemental composition and heat treatment parameters on the γ′ depletion zone thickness, providing valuable insights for thermal barrier coating bond coat design. However, such post-hoc interpretation techniques only yield qualitative rankings that lack physical grounding. They fail to establish a quantitative mapping between physicochemical features and material properties, thus falling short of offering deep physical interpretability.

To extract and reuse design pathways from complex models, projection-based dimensionality reduction techniques are commonly applied to visualize feature-property relationships in two dimensions. Chen *et al.*[42] used the Principal Component Analysis (PCA)[43] to project multiple variables into two-dimensional space, successfully visualizing the stability boundary of solid-solution phase in a five-component high-entropy alloy (HEA) system. Their analysis identified melting point and mixing enthalpy as the primary factors determining phase stability. Srinivasan *et*

al.[44] employed the Isomap algorithm[45] to map high-dimensional descriptors of the Co_3(Al, X) system into a low-dimensional space, generating a 2D point-line graph where elements are represented as nodes and descriptor similarities as edges, thereby intuitively identifying candidate compositions with high binding energy, high melting point, and coherence with the Co-rich face-centered cubic matrix. Tian *et al.*[32] applied Uniform Manifold Approximation and Projection (UMAP)[46] to reduce the 768-dimensional vectors from steel-specialized large language model (SteelBERT) into a two-dimensional space while preserving local topological structures. The resulting "knowledge map of steel" was distinguished by keywords such as fatigue, irradiation and welding, which intuitively revealed latent research hotspots and trends within the literature. These methods share the common objective of mapping high-dimensional data into a two-dimensional space to generate clear and intuitive visualizations that facilitate the analysis and structural understanding of complex data. However, the relationship between the post-reduction features and the original features remains obscure, making it impossible to analyze the relationship between the features before and after dimensionality reduction and the impact on material properties.

In summary, current approaches still exhibit significant limitations in providing model interpretability and the visualization of material design. Previous studies have shown that algorithms such as symbolic regression and linear regression can establish explicit mathematical expressions linking material features to properties, presenting a viable path to overcome these shortcomings. Zhao *et al.*[47] employed symbolic regression to establish an explicit mathematical expression linking alloy composition and thermal conditions to high-temperature specific yield strength, providing a closed-loop framework for the inverse design of refractory high-entropy alloys. Xue *et al.*[48] successfully developed a simple and interpretable polynomial regression expression relating the phase transformation temperature of NiTi-based shape memory alloys to crystal structure parameters. Such approaches extract compact, physically informative expressions from materials data to guide design. However, the predictive accuracy is limited due to the oversimplification inherent in representing material behavior with only a small set of features.

To address these challenges, this work proposes a novel Two-dimensional Symbolic Feature Generation (2D-SFG) method via genetic programming-based symbolic

regression, enabling visual and interpretable materials design. By combining original features with mathematical operators, 2D-SFG constructs explicit mathematical expressions as new features, which are then used to generate two-dimensional classification probability maps or property contour plots. This supports efficient and intuitive materials design. Applied to various classification and regression tasks, the method simultaneously improves predictive accuracy, enhances interpretability, and enables design visualization, thereby establishing a new strategy for interpretable and visual materials design.

MATERIALS AND METHODS

This work introduces a novel feature construction methodology designed to simultaneously enhance the predictive performance of machine learning models and provide explicit, interpretable feature representations. Mathematically, for a given ML problem formulated as **target** = $f(x_1, x_2, \ldots, x_n)$, two new features are generated through symbolic regression: $X_1 = F_1(x_1, x_2, \ldots, x_n)$ and $X_2 = F_2(x_1, x_2, \ldots, x_n)$. This transforms the original modeling task into **target** = $g(X_1, X_2)$, with the key constraint that the model based on these constructed features achieves superior accuracy or lower error compared to the initial model. Notably, F_1 and F_2 represent explicit mathematical relationships rather than uninterpretable black-box models. The overall workflow is illustrated in Figure 1.

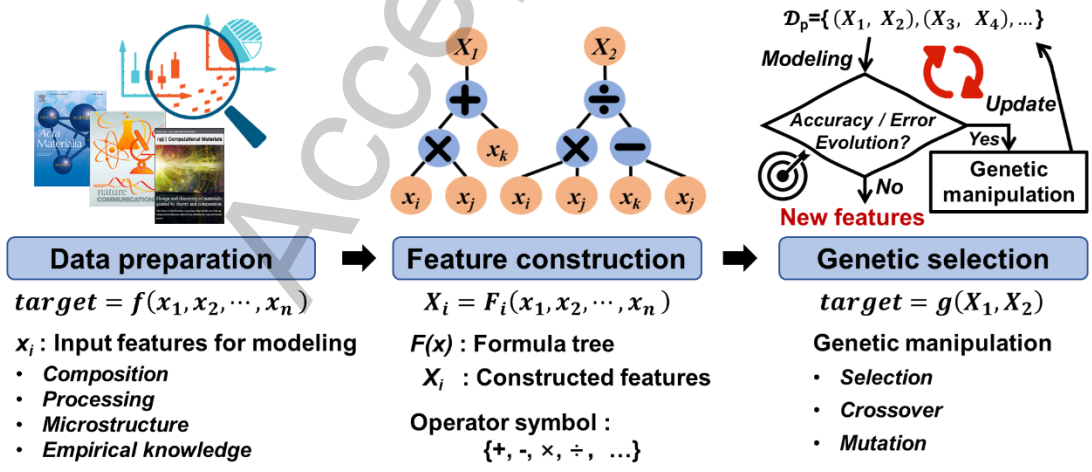


Figure 1. Two-dimensional symbolic feature generation method. A novel strategy of feature construction for two-dimensional symbol feature combining symbolic regression and genetic algorithm.

First, features relevant to the modeling problem, such as composition, processing conditions, microstructure and empirical knowledge, were collected to form an initial feature set. It is noteworthy that this set may comprise either raw, unscreened features or a pre-selected subset of key attributes identified in earlier research. Symbolic regression was then applied to construct new features in the form of explicit mathematical expression trees derived from the original features. This construction process constitutes the core of the proposed methodology, aiming to synthesize critical information from the initial feature set through interpretable mathematical expressions while eliminating redundancy and irrelevant details.

However, since feature construction typically explores an expansive combinatorial space, an exhaustive search over all possible expressions is computationally infeasible. Therefore, an evolutionary optimization approach was introduced to address this issue. Newly constructed features were paired to form a "feature-pair set", with a population size set to 500. Each feature pair served as input for model training and the model was evaluated based on a fitness metric (e.g., accuracy or error) to retain high-performing feature-pairs. Subsequently, crossover and mutation operations were applied to generate new feature-pair sets over 100 generations, progressively improving feature quality and model performance. In the Genetic Algorithm, Accuracy serves as the fitness function for classification tasks, while Mean Absolute Percentage Error or Root Mean Square Error is employed for regression tasks. Concurrently, a hard constraint limiting the maximum number of nodes to fewer than 17 is imposed to balance model accuracy with formula interpretability. The pseudo-code of the proposed 2D-SFG method is summarized in Table 1.

Table 1. Pseudocode of 2D-SFG method

Step	Methods
1	Let $\mathcal{D}_o$:= {+, -, ×, ÷, …} and $\mathcal{D}_f := \{x_1, x_2, \ldots, x_n\}$ represent the operator symbol set and feature set, respectively.
2	Generate N formula trees $\{X\}$ with the elements randomly selected from the two sets in **Step 1**, e.g., $X = x_i x_k - x_j, x_i / x_j, \ldots$.
3	A set of feature pairs, denoted as $\mathcal{D}_p$, was generated by sampling 2N times from {X} with replacement, i.e., $\mathcal{D}_p = \{(X_1, X_2), (X_3, X_4), (X_1, X_3), \ldots\}$, in

which ***X*** represents the formula tree in **Step 2.**

4 For each pair in $\mathcal{D}_p$, build a model with the 2 features as inputs and evaluate the model.

5 ***If*** Stop condition **not** met:
Update formula trees by genetic algorithm operations (***Selection, Crossover, Mutation***).
Repeat from **Step 4.**
Else:
Return the optimal feature pair.

End

To mitigate symbolic overfitting and expression bloat, the following control mechanisms were implemented:

- Operator & Constant Restrictions: The search space was limited to basic arithmetic operators {+, −, ×, /} and a fixed set of constants {0, 1, 2}. This avoids the complexity of optimizing floating-point coefficients and ensures analytical simplicity.
- Structural Hard Constraints: A strict limit of maximum 17 nodes per symbolic tree was enforced. Any genetic operation violating this constraint was rejected.
- No Explicit Penalties or Pruning: Unlike methods using description length penalties, our approach relies on hard constraints to naturally limit complexity. Additionally, redundant sub-expressions (e.g., x+0) were simplified during generation, eliminating the need for post-hoc pruning.
- Numerical Stability: Protected division was employed to handle near-zero denominators, preventing numerical errors during fitness evaluation.
- Dimensional Flexibility: No strict dimensional consistency was enforced, allowing the discovery of empirical dimensionless groups common in materials science.
- Independent Testing Set: A completely independent test set was established. The discrepancy in predictive performance between the training set and this independent testing set was utilized to evaluate the risk of overfitting.

RESULTS AND DISCUSSION

Material visualization classification model

Feature dimensionality reduction techniques are commonly employed to map high-dimensional data into a lower-dimensional space for more intuitive visualization of sample distributions across classes in classification tasks, such as principal component analysis. These methods primarily retain directions of maximum variance to highlight inter-class differences. However, high variance does not necessarily correspond to discriminative or physically meaningful information. For instance, noise may exhibit greater variance than critical target features. In such cases, reducing the input dimensionality can lead to the loss of important information, thereby damaging the classification performance of the model[49]. The proposed 2D-SFG method not only reduces the input dimensionality by constructing new features but also incorporates nonlinear combinations among features, thereby effectively preserving the critical information from the original data while enhancing model visualization and improving predictive accuracy. To evaluate its effectiveness, we compared its performance with existing results in perovskite structural categorization and high-entropy alloy phase classification, as reported in high-impact journals including ***Nature Communications*** and ***Acta Materialia***.

Balachandran *et al.*[30] developed a support vector machine-based binary classification model using a dataset of 167 perovskite structures and five input features, achieving an average accuracy of 85.6%. Based on high-entropy alloy data gathered from references[50-52], a phase classification model was built using ten influential features selected via a feature screening method, resulting in a prediction accuracy of 84.5%. The material features employed in the two classification tasks are presented in Table 2. Comprehensive details can be found in the **Supplementary Materials**.

Table 2. Features used in classification tasks

Task	Features	
Perovskites	f_0 **:** Tolerance factor	f_1 **:** Valence electron number
	f_2 **:** Martynov-Batsanov electronegativity	f_3 **:** Ideal bond lengths

High-entropy alloys	f_4 **:** Mendeleev number	
	ΔH_{mix} **:** Mixing enthalpy	Xc **:** Chemical bond mismatch
	$\overline{MAC}$ **:** Mean mass attenuation coefficient	$\overline{EA}$ **:** Mean electron affinity
	δCE **:** Cohesive energy deviation	$\overline{RM}$ **:** Mean rigidity modulus
	$\overline{MN}$ **:** Mean mendeleev number	δMR **:** Metal radius deviation
	δAV **:** Atomic volume deviation	$\overline{LC}$ **:** Mean lattice constants a

Figures 2A and 2B illustrate the iterative feature generation process for two classification cases. As iterations proceed, the model accuracy gradually increases until it converges to a stable value. This trend indicates that the two new features generated by the 2D-SFG method incorporate more meaningful information to improve model performance. The explicit mathematical relationships of the evolved features after 100 iterations are shown in Equations (1) to (4).

Perovskites:

$$X_{C1} = f_3 - f_3^2 - f_2 \times \frac{f_3}{f_1} \quad (1)$$

$$X_{C2} = f_4 - f_2 \quad (2)$$

High-entropy alloys:

$$X_{C3} = \delta MR \times \delta AV \times \left(2 \times \overline{MAC} \times \frac{\overline{EA}}{\overline{LC}} + \overline{RM} + \delta AV \right) \quad (3)$$

$$X_{C4} = \overline{MN} - Xc + \frac{\delta AV}{\delta CE} + \frac{\overline{MN}}{\overline{EA}} \quad (4)$$

Where X_{C1}, X_{C2}, X_{C3} and X_{C4} denote the constructed symbolic features. Definitions for all other parameters can be found in Table 2.

A support vector classifier was established using the constructed features, and the corresponding classification probability maps are shown in Figures 2C and 2D. On the one hand, these maps provide an intuitive representation of the machine learning model, in which the red arrows indicate the direction of increasing probability for the target

property, forming a "navigation map" for alloy design. Based on the value range associated with the positive class, the composition design window for ferroelectric perovskites without secondary phases and single solid-solution high-entropy alloys can be determined through either forward screening or inverse calculation. On the other hand, underlying patterns in the data can be revealed by analyzing the influence of the constructed features on the target property, particularly the transition boundaries. It offers possibilities for extracting physical interpretability and discovering new knowledge.

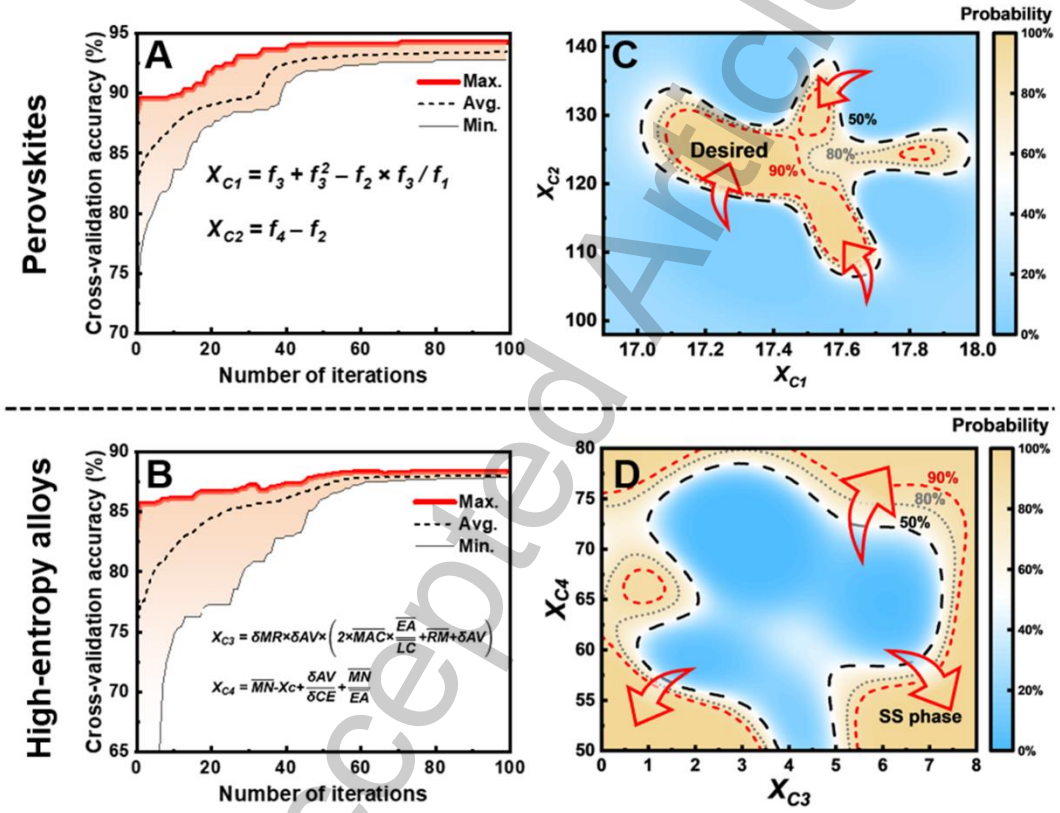


Figure 2. 2D-SFG method applied to classification tasks. (A and B) Feature iteration generation process, the red curve indicates the change of optimal performance in the population; (C and D) Classification probability maps, the red arrows indicate the direction of increasing probability for the target property.

Material visualization regression model

In regression tasks, the "curse of dimensionality" also poses a significant challenge. Although reducing the dimension of the input variables can visualize the relationship between the input and the output, it is impossible to obtain the relationship between the features before and after dimensionality reduction. The proposed 2D-SFG method can

not only preserve meaningful information and eliminate redundancies during dimensionality reduction but also provide explicit mathematical expressions that clarify the connections between original and transformed features, as well as their influence on material properties. This approach significantly enhances model interpretability and offers valuable support for in-depth investigation into how key features affect performance mechanisms in materials science. To further assess the effectiveness of our approach, we selected studies on machine learning-assisted alloy design published in high-impact journals including ***Nature Communications*** and ***Acta Materialia*** for comparative validation.

Xue *et al.*[53] established a support vector regression prediction model with 6 features as input to design the composition of the NiTi-based shape memory alloy with low thermal hysteresis, achieving a prediction error of 2.9 K. The authors[9] aimed to optimize both hardness and electrical conductivity in precipitation strengthened copper alloys. Five key features influencing alloy hardness and six features affecting electrical conductivity were identified through feature selection. By constructing corresponding support vector regression prediction models, the prediction errors achieve 5.9% and 9.7%, respectively. The material features used in these two regression tasks are summarized in Table 3, with additional details provided in the **Supplementary Materials**.

Table 3. Features used in regression tasks

Task	Features	
Thermal hysteresis	***en :*** Pauling electronegativity	***cs :*** Pettifor chemical scale
	ven : Valence electron number	***arc :*** Clementi's atomic radii
	dor : Waber-Cromer pseudopotential radii	***mr :*** Metallic radius
Vickers hardness	***M.S3 :*** Covalent radii	***M.S12 :*** Lattice constants c
	V.A8 : Mass attenuation coefficient variance	***V.S7 :*** Atom volume variance

	O12 : Solubility at room temperature	
Electrical conductivity	***M.A10 :*** Mass attenuation coefficient	***M.S6 :*** Core electron distance
	M.E4 : Absolute electronegativity	***V.E6 :*** Second ionization energy
	M.C4 : Melting enthalpy	***HV :*** Vickers hardness

Figures 3A to 3C illustrate the iterative feature generation process for two regression cases. As the number of genetic iterations increases, the 10-fold cross-validation error progressively decreases, indicating that the two constructed features incorporate increasingly meaningful information, thereby contributing to model improvement. Notably, the error for the hardness regression task continued to decrease beyond 100 iterations. However, further increasing generations results in overly complex expressions. This reflects the need for more sophisticated mathematical representations to accurately model the underlying strengthening mechanisms and integrate the substantial information embedded within the original features. To balance model accuracy with expression complexity, the evolutionary process was terminated at 100 generations. The final mathematical expressions of the evolved features after 100 generations are provided in Equations (5) to (10).

ΔT:

$$X_{R1} = (arc+dor)\times cs-dor-ven \quad (5)$$

$$X_{R2} = \frac{en}{mr\times(2cs+arc+mr+en+dor)} \quad (6)$$

HV:

$$X_{R3} = \frac{M.S3 + O12}{V.A8} + V.S7-O12-\frac{V.A8}{M.S12} \quad (7)$$

$$X_{R4} = M.S3\times V.S7+M.S3-M.S12-\frac{V.S7-1}{V.A8\times O12-M.S3} \quad (8)$$

EC:

$$X_{R5} = M.A10\times M.S6-M.A10 \quad (9)$$

$$X_{R6} = \frac{M.A10}{M.S6\times HV} + 3\times M.E4 \quad (10)$$

Where X_{R1}, X_{R2}, X_{R3}, X_{R4}, X_{R5} and X_{R6} denote the constructed symbolic features. Definitions for all other parameters can be found in Table 3.

Support vector regression models were established using the constructed features as input to predict alloy properties across different values, as visualized in Figures 3D to 3F. The red arrows in each contour plot indicate the gradient variation of the target property, serving as guidance for efficiently approaching the desired property objectives. For instance, thermal hysteresis is minimized when features X_{R1} and X_{R2} approach values around 145 and 4.2×10^{-5}, respectively. Accordingly, NiTi-based shape memory alloys with low thermal hysteresis can be designed by forward screening or inverse calculation of compositions that satisfy this specific feature range Similarly, higher alloy hardness is achieved near $X_{R3} \approx -4$ and $X_{R4} \approx -246$, while superior electrical conductivity is attained around $X_{R5} \approx 12.5$ and $X_{R6} \approx 14.2$. These identified value ranges offer practical and interpretable design guidelines for developing copper alloys with an improved balance of hardness and conductivity.

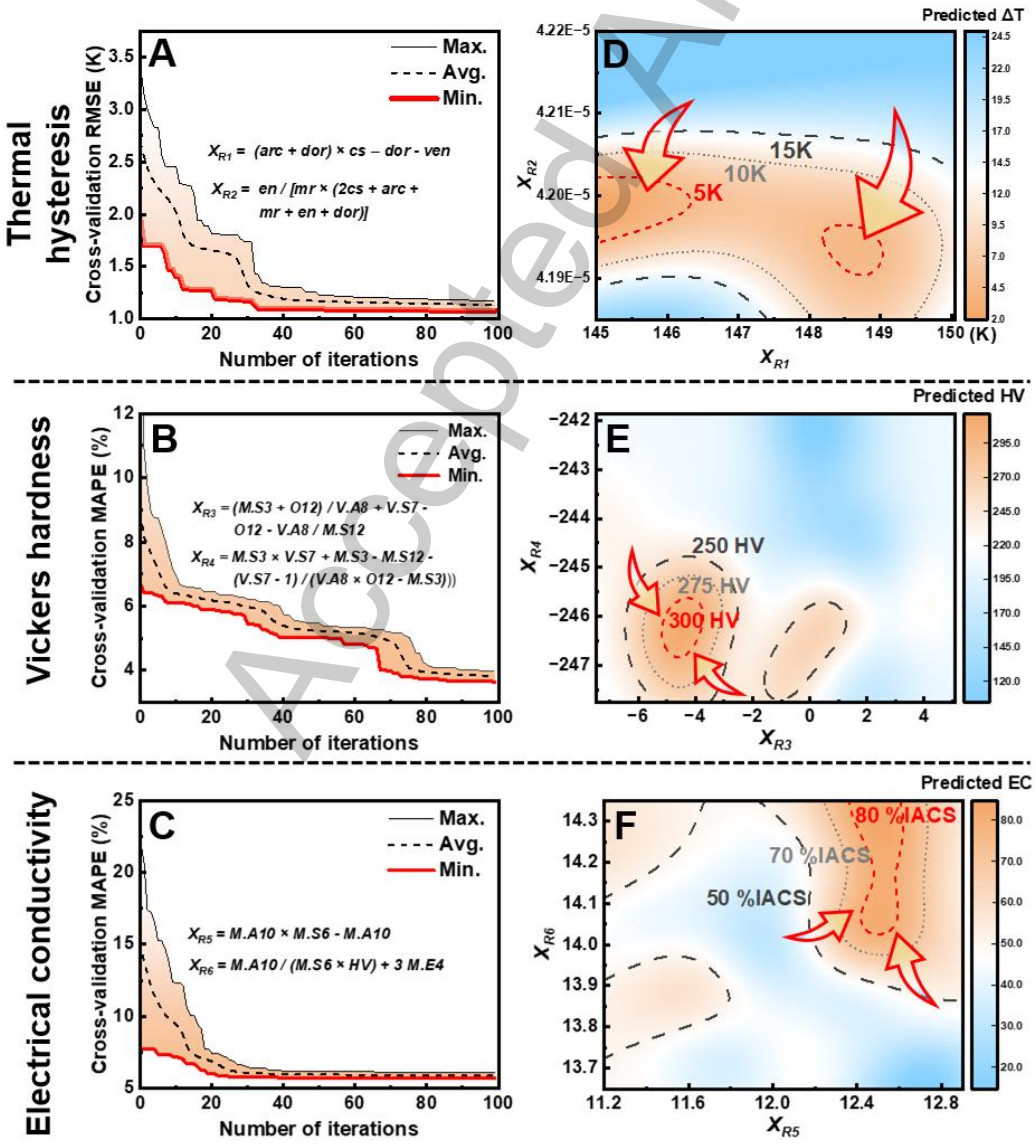


Figure 3. 2D-SFG method applied to regression tasks. (A to C) Feature iteration generation process, the red curve indicates the change of optimal performance in the population; (D to F) Contour map of alloy properties, the red arrows indicate the gradient variation of the target property.

Material visualization intelligent design

As illustrated in Figures 2 and 3, the 2D-SFG method effectively visualizes phase boundaries and property contours, enabling intuitive identification of target regions and variation trends. A direct correspondence between composition design windows and target properties was established by mapping candidate alloy compositions into two-dimensional symbolic features. In this work, we employed the 2D-SFG method to construct visualization models that compute the symbolic features for unknown alloy compositions. These models were subsequently used to evaluate and predict target properties, successfully screening refractory high-entropy alloys with single solid-solution phases as well as advanced copper alloy compositions with superior performance. In the V-Nb-Ti-W-Zr (quinary alloy) high-entropy alloy system, a genetic algorithm was used to define a composition space with 1at.% increments. The ten-dimensional features listed in Table 2 were computed for each composition and projected into a two-dimensional visualization space using Equations (3) and (4), thereby identifying candidate compositions with a high probability of forming solid-solution phases, as summarized in Table 4.

Table 4. Composition of high-entropy alloys

Alloy (at. %)	V	Nb	Ti	W	Zr	Mo	Hf	Ta
HEA-1	30	29	8	12	10	9	1	1
HEA-2	30	27	18	13	10	1	1	0

Figure 4A displays the classification probability and evolutionary path of alloys across two generations, showing a clear design pathway toward higher solid-solution probability. The selected alloys were synthesized using vacuum arc melting and characterized by X-ray diffraction. As shown in Figure 4B, all as-cast alloys exhibit a single BCC (body-centered cubic) solid-solution phase, consistent with the predictions

of the 2D-SFG method. Figure 4C compares the predictions of different visualization modeling approaches. While most methods struggle to accurately identify the phase structure of alloys with low solid-solution probability shown in Figure 4A, only the 2D-SFG method correctly predicts the phase structure for both alloys. These results underscore the strong generalization capability and classification accuracy of the proposed approach in identifying phase in high-entropy alloys.

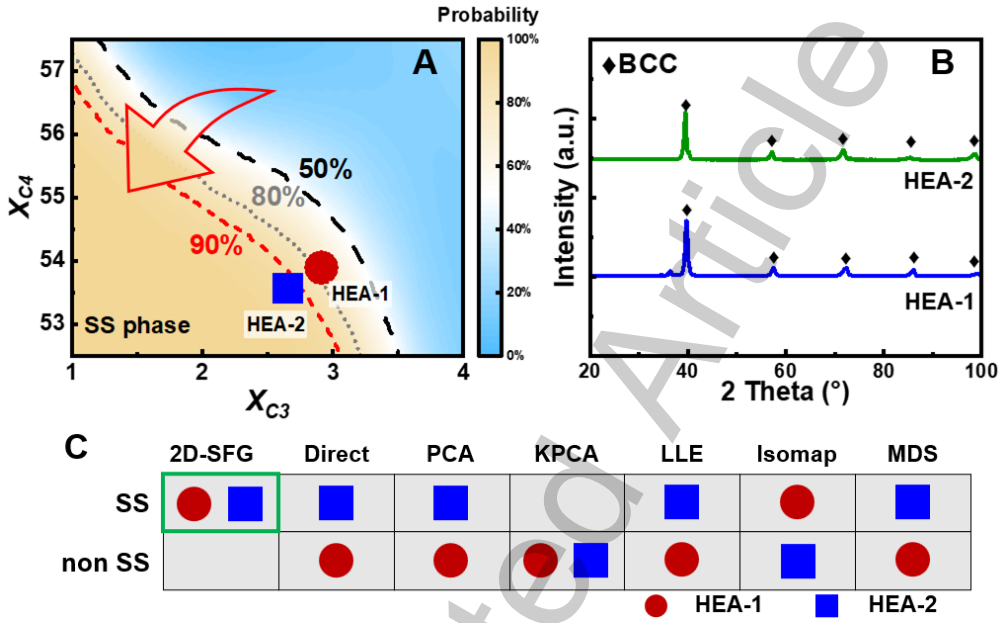


Figure 4. 2D-SFG method for high entropy alloy design. (A) Local classification probability map; (B) XRD characterization results; (C) Prediction results of phase by different methods.

Figure 5 illustrates the application of the constructed hardness and conductivity visualization models to the unexplored composition space within the Cu-Ni-Si system. A promising copper alloy composition, Cu-1.87Ni-0.55Si-0.47Co-0.11Mg-0.07Zr-0.21Zn, was identified on the performance Pareto front as exhibiting a favorable high-hardness and medium-conductivity profile. The designed alloy was synthesized via vacuum induction melting, achieving a peak hardness of 275 HV and a corresponding electrical conductivity of 47.2 %IACS (percent International Annealed Copper Standard). As summarized in Table 5, the prediction performance of different design methods was compared. The 2D-SFG model demonstrated superior generalization capability, with all errors between experimental measurements and model predictions remaining below 7%, confirming its effectiveness in supporting the rational design of

high-performance alloys.

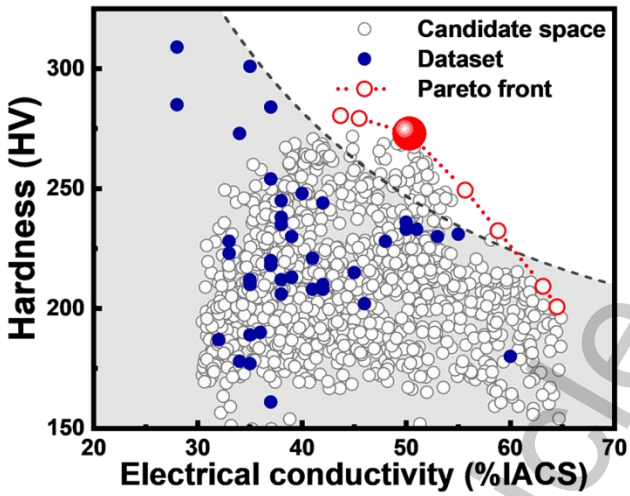


Figure 5. The properties of the designed alloys within the Cu-Ni-Si system. Blue solid dots represent the collected experimental dataset, while white hollow dots represent the explored composition design space. The dashed line indicates the identified Pareto front, with the red data point marking the optimal alloy composition selected through the 2D-SFG guided design approach.

Table 5. prediction results of properties by different methods

Methods	2D-SFG	Direct	PCA	KPCA	LLE	Isomap	MDS
HV_{pred} (HV)	273	268	209	233	213	215	216
Percentage error	**0.7%**	2.5%	24.1%	15.3%	22.5%	21.8%	21.5%
EC_{pred} (%IACS)	50.3	43.8	54.2	46.6	58.1	37.6	43.7
Percentage error	**6.6%**	7.2%	14.8%	1.3%	23.1%	20.3%	7.4%

Material visualization intelligent design model performance

To evaluate the performance of the 2D-SFG method in classification tasks, a comprehensive assessment was conducted using multiple metrics, including accuracy, precision, recall, F1-score and AUC, as shown in Figures 6A and 6B. The model constructed with the 2D-SFG features achieved accuracy and precision rates above 90% on both the training and test sets, with only minor differences between the two,

indicating no significant overfitting. The F1-score remained consistently above 0.9, reflecting a well-balanced trade-off between precision and recall. Furthermore, the AUC values for both training and test sets were significantly higher than the 0.9 threshold, demonstrating the model's strong discriminative ability for classification boundaries and its reliability for practical applications.

A comparative analysis was conducted between direct modeling and the 2D-SFG approach, as illustrated in Figure 6C. The direct modeling method utilized the initial feature set from Table 2 as input, while the 2D-SFG approach employed the new features represented by Equations (1) to (4). It should be noted that both methods were evaluated using the same training and test datasets, and both underwent hyperparameter optimization. The 2D-SFG model achieved prediction accuracies of 94.2% for perovskite structure classification and 88.4% for high-entropy alloy phase classification, surpassing the 85.6% and 84.5% obtained by the initial feature-based model. These results demonstrate that the 2D-SFG method effectively extracts meaningful information from the original features while reducing interference from redundant or irrelevant data.

From the perspective of feature reduction, the proposed 2D-SFG method can also be regarded as a dimensionality reduction technique. To evaluate its effectiveness, this work compared it against several common dimensionality reduction methods, including PCA, KPCA (Kernel Principal Component Analysis)[54], LLE (Locally Linear Embedding)[55], Isomap and MDS (Multidimensional Scaling)[56]. All methods were configured to reduce the original features to two dimensions before model construction. As shown in Figure 6D, models using features from conventional reduction methods exhibited varying degrees of accuracy degradation, indicating the loss of critical information during the reduction process. In contrast, the 2D-SFG approach not only preserved essential information but also enhanced predictive accuracy, demonstrating its superior applicability for classification tasks.

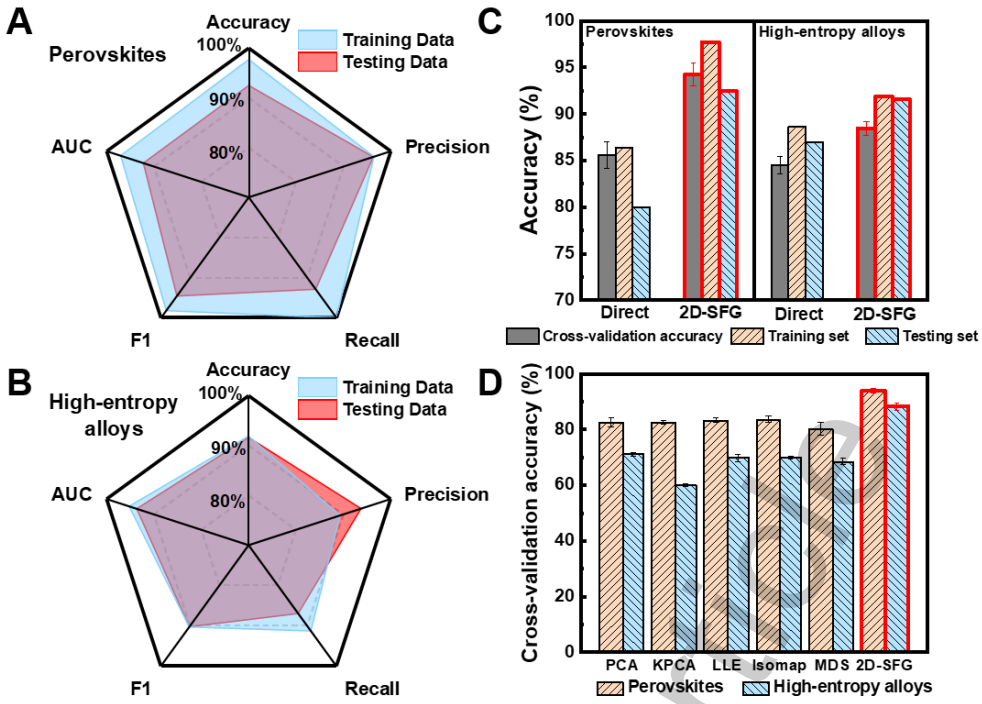


Figure 6. Performance of 2D-SFG methods on classification problems. (A and B) Evaluation indexes of classification model; (C) Comparison of modeling accuracy between direct modeling and symbolic feature generation; (D) Multiple dimensionality reduction methods VS 2D-SFG method. Error bars indicate the standard deviation obtained from 10-fold cross-validation.

For the regression tasks, a similar comparative analysis was performed among direct modeling, 2D-SFG modeling and models built after applying common dimensionality reduction methods, evaluated using 10-fold cross-validation errors. The evaluation was based on root mean square error (RMSE) and percentage error metrics. As shown in Figure 7, models using conventional dimensionality reduction showed either no significant change or a slight increase in error, indicating substantial limitations of these methods for regression problems. In contrast, the 2D-SFG approach achieved prediction errors of 1.1 K, 3.7% and 6.2% for shape memory alloy thermal hysteresis, copper alloy hardness and electrical conductivity, respectively. These results are lower than the corresponding errors of 2.9 K, 5.9% and 9.7% obtained using the original feature set, demonstrating the strong applicability of the 2D-SFG method for regression tasks.

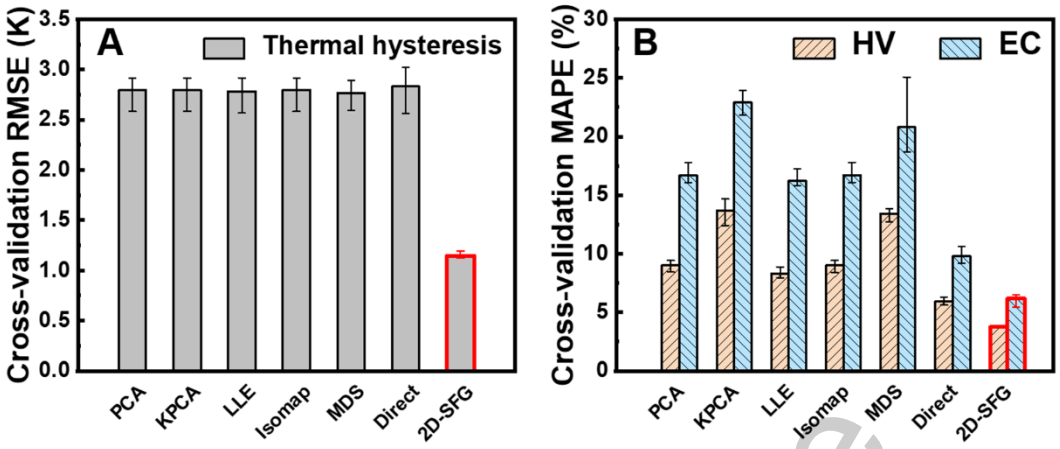


Figure 7. Multiple dimensionality reduction methods VS SFG method. (A) Thermal hysteresis for shape memory alloy; (B) Hardness and electric conductivity for copper alloy. Error bars indicate the standard deviation obtained from 10-fold cross-validation.

Although our primary analysis and discussion focused on the Support Vector Machine algorithm due to its robustness in small-sample materials science problems, the core advantage of 2D-SFG, as a feature generation method for physics-informed knowledge discovery, lies in its ability to generate mathematical descriptors that explicitly encode physical mechanisms. This alleviates the burden on black-box models to implicitly extract features, thereby enhancing both predictive accuracy and generalization capability. To verify whether these performance gains are algorithm-agnostic, we conducted additional comparative studies using Random Forest (RF) and Gradient Boosting (GB) algorithms. The results, presented in Supplementary Table S3, demonstrate that models trained on symbolic features generated by 2D-SFG consistently outperform those based on origin feature inputs across all tested algorithms. These findings indicate that 2D-SFG serves as a versatile and seamlessly compatible module capable of integrating into diverse machine learning workflows. Extending the 2D-SFG framework to guide more complex architectures, such as Deep Neural Networks, will be a primary focus of our future work.

Generated feature importance and interpretability analysis

The occurrence frequency of each original feature in the explicit expressions generated at the 100 iterations was counted to assess feature importance, as illustrated in Figure 8. A higher frequency indicates greater importance of the corresponding feature. For instance, feature f_0 does not appear in Equation (2), suggesting that its information is

either redundant with the other four features or has limited influence on the target property. Consequently, it was discarded in later iterations, with an overall frequency of less than 20. In the case of copper alloy hardness, all features exhibited occurrence frequencies above 100, reflecting their substantial individual importance and joint contribution to the target property. This also explains why the final explicit expression for hardness is more mathematically complex. In contrast, the electrical conductivity of precipitation strengthened copper alloys is governed by a relatively simpler mechanism compared to hardness. As a result, meaningful information could be extracted from the original features using simpler mathematical operations, leading to a less complex symbolic expression for conductivity.

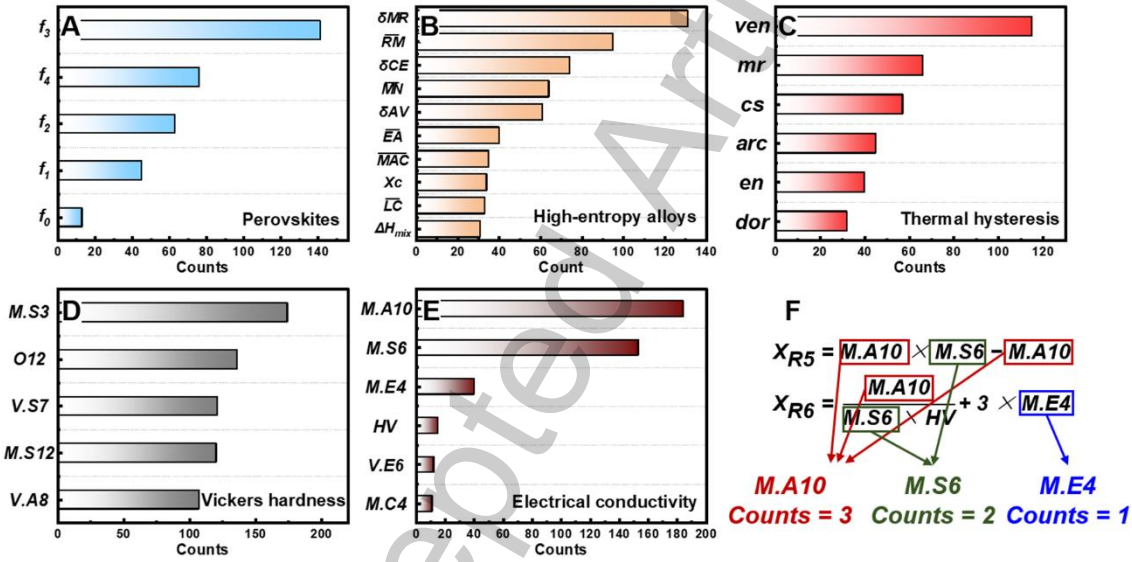


Figure 8. Feature importance ranking. (A) Structure classification task for perovskites; (B) Phase classification task for high-entropy alloys; (C) Thermal hysteresis regression task for shape memory alloys; (D and E) Comprehensive property regression task for copper alloys. (F) Example of counting rules.

This work successfully constructed new features for various material classification and regression problems through the 2D-SFG method. These features not only significantly enhance the predictive performance of machine learning models but also maintain explicit quantitative relationships with the original features, thereby improving physical interpretability for specific materials research problems. The 2D-SFG method is capable of constructing meaningful nonlinear combinations that capture complex physical interactions. By analyzing the specific mathematical forms of the generated

symbolic features, such as products, ratios, and linear corrections, which can elucidate how feature co-occurrence influences target properties.

In the classification tasks, the new feature X_{C1} couples the bonding strength driven by electronegativity (f_2) with the valence electron number (f_1). This formulation embodies the electronic shielding effect, where a lower valence electron count amplifies the impact of charge rearrangement, induced by electronegativity differences, on lattice distortion. The formula demonstrates that the formation of the perovskite requires not only geometric matching but also that the polarity strength of chemical bonds be weighted and corrected by the electron concentration[57]. The new feature X_{C3} employs a multiplicative form to quantify the nonlinear geometric distortion driving force arising from the dual deviations in metal radius (δMR) and atomic volume (δAV). The formation of high-entropy alloy solid solutions requires not only minimal geometric distortion but also that the material's intrinsic electronic properties and mechanical stiffness possess the capacity to effectively buffer and accommodate the inevitable lattice strains[58,59].

In the regression tasks, the original features atomic radius (***arc***) and pseudopotential radius (***dor***) associated with atomic size mismatch frequently co-occur. An increase in their values can raise the distortion energy barrier, thereby increasing thermal hysteresis. The new feature X_{R1} multiplied atomic size-related features with a chemical scaling parameter to capture the synergistic effect between atomic size mismatch and the chemical bonding environment, while subtracting the valence electron number (***ven***) to correct for the shielding effect on lattice distortion in high electron concentration systems. This formulation comprehensively reflects the combined influence on thermal hysteresis[60]. Furthermore, the new feature X_{R2} is expressed as a ratio of electronegativity to multi-scale structural parameters. Physically, a larger denominator indicates stronger lattice distortion and greater resistance to the phase transformation driving force, while a larger numerator reflects more intense electron transfer and a higher phase transformation energy barrier. Consequently, a smaller value of X_{R2} facilitates easier phase transformation, corresponding to lower thermal hysteresis[61,62]. The new feature X_{R5} is formulated as the product of the mass attenuation coefficient and the extranuclear electron distance. This construction captures the coupling between electron scattering efficiency and atomic size effects, where the extranuclear electron

distance governs electron cloud distribution. Enhanced electron cloud overlap, driven by this interaction, leads to significantly improved conductivity. Additionally, the new feature X_{R6} incorporates measured Vickers hardness and absolute electronegativity. This combination accounts for the influences of lattice distortion and dislocation density. This formulation aligns well with established understanding in metallic systems, where electrical conductivity is collectively determined by the interplay of lattice defects, solute atoms, and polarization effects[63-65].

The 2D-SFG method demonstrates a unique capability to automatically discover nonlinear descriptors that are difficult to predefine empirically. Although the resulting feature expressions may appear mathematically complex, each term corresponds to established physical models or empirical relationships in materials science. Furthermore, these explicit mathematical expressions effectively circumvent the lack of physical insight inherent in black-box models. However, it is important to acknowledge certain limitations of the generated features: the physical universality of these expressions requires validation across broader material systems, and certain high-order interaction terms might potentially introduce singularities. Future work should focus on enhancing the physical consistency and stability of the feature engineering process.

Analysis of the reuse effect of generated features

To further evaluate whether the symbolic feature generation method successfully integrates and learns universal knowledge, the authors collected a dataset of aluminum alloys with solution and aging treatments. This dataset includes composition, processing parameters, hardness and electrical conductivity. The initial features listed in Table 3 were calculated, with the solid solubility of elements in aluminum being used as a substitute for their solubility in copper. As shown in Table 6, using the aluminum alloy data to calculate features X_{R3} to X_{R6} and retrain the model (Method A) yielded better predictive performance than the model trained with the original aluminum alloy features (Method B). For hardness prediction, Method A achieved cross-validation, training, and test errors of 9.9%, 7.5% and 8.4%, respectively, representing an average reduction of 2.0% compared to Method B (10.9%, 9.6%, 11.3%). Similarly, for electrical conductivity, the errors decreased from 3.8%, 3.6% and 5.3% to 2.9%, 1.1% and 2.4%, with an average improvement of 2.1%. When a random forest classifier was

used instead, the two-dimensional symbolic features still provided an average reduction in hardness error of 1.8 % and in conductivity error of 2.2 %, confirming that the advantage stems from the features themselves rather than algorithmic coincidence.

Table 6. Prediction results of models under different methods

Methods	Hardness			Electrical conductivity		
	CV	Training	Testing	CV	Training	Testing
Method A (SVR)	**9.9%**	**7.5%**	**8.4%**	**2.9%**	**1.1%**	**2.4%**
Method B (SVR)	10.9%	9.6%	11.3%	3.8%	3.6%	5.3%
Method A (RF)	12.3%	10.8%	13.3%	3.2%	3.1%	7.6%
Method B (RF)	14.5%	11.6%	15.7%	4.3%	5.7%	10.5%

Note: The bolded values highlight the superior results achieved by the proposed method relative to the other benchmarks.

The distributional differences in feature values between copper and aluminum alloys pose a challenge of out-of-distribution generalization when directly applying copper alloy models to aluminum alloy data. Although Model B was retrained using aluminum alloy data, the original features may still fail to capture the complex nonlinear relationships governing material properties. In contrast, the symbolically generated features constructed from copper alloy data preserve underlying mappings to alloy performance, as evidenced by the superior predictive performance of the aluminum alloy regression model (Model A) built with these features. The strong performance of Model A confirms both the applicability and effectiveness of the symbolically generated features, which not only retain key physical information learned through the 2D-SFG method but also adapt effectively to aluminum alloy data through refitting. The demonstrated predictive capability across both copper and aluminum alloy systems validates the transferability of the 2D-SFG method, offering a novel approach for cross-system materials modeling.

CONCLUSIONS

This study aimed to overcome the dual challenges of limited model interpretability and the difficulty in extracting reusable design routes in materials informatics by developing a novel Two-Dimensional Symbolic Feature Generation method. Our findings

demonstrate that integrating symbolic regression with genetic algorithms not only significantly enhances predictive accuracy across diverse material systems but also yields compact mathematical formulas that align with classical physical models, thereby ensuring visualization and physically interpretable dimensionality reduction. Furthermore, the generated 2D feature maps provide intuitive visualizations of phase boundaries and property contours, effectively guiding the successful design of high-performance alloys without relying on black-box predictions. This interpretable framework lays the foundation for a fundamental shift in materials design methodology, moving from 'black-box' exploration to 'white-box' rational design by exploring the relationship between data and fundamental physical mechanisms.

DECLARATIONS

Authors' contributions

W.Y and H.T.Z contributed equally to this work. W.Y and H.T.Z carried out the majority of the modeling and experimental work. Z.L. and J.H. contributed to the discussion and data analysis. C.B.C. and Y.X.G conducted data collection. H.D.F. and J.X.X. designed modeling approaches and supervised the research. All authors interpreted the results and contributed to writing the paper.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Financial support and sponsorship

This work was supported by the National Major Science and Technology Projects of China (2024ZD0608100), National Natural Science Foundation for Distinguished Young Scholars of China (No. 52425409), National Natural Science Foundation of China (Nos. 52404387, 52374379, and U24A2029), Xiaomi Young Scholars Program, and Interdisciplinary Research Project for Young Teachers of USTB (Fundamental Research Funds for the Central Universities) (No. FRF-IDRY-23-002).

Conflicts of interest

Huadong Fu is a Guest Editor of Special Topic "AI-Driven Design and Intelligent Manufacturing of Advanced Copper Alloy" of the journal *Journal of Materials*

Informatics, but was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, and decision making, while the other authors have declared that they have no conflicts of interest.

Additional information

Supplementary Materials to this article can be found online or from the author.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Rahman, A.; Hossain, M. S.; Siddique, A. B. Review: Machine learning approaches for diverse alloy systems. *J. Mater. Sci.* **2025**, 60, 12189-12221. DOI: 10.1007/s10853-025-11154-4.

2. Hu, M.W. *et al.* Recent applications of machine learning in alloy design: A review. *Mater. Sci. Eng. R.* **2023**, 155, 100746. DOI: 10.1016/j.mser.2023.100746.

3. Kumar, A.; Mukhopadhyay, N. K.; Yadav, T. P. Recent progresses on high entropy alloy development using machine learning: A review. *Comput. Mater. Today.* **2025**, 8, 100038. DOI: 10.1016/j.commt.2025.100038.

4. Cheng, M. Y. *et al.* Artificial intelligence-driven approaches for materials design and discovery. *Nat. Mater*. **2026**, 25, 174-190. DOI: 10.1038/s41563-025-02403-7.

5. Hart, G. L. W.; Mueller, T.; Toher, C.; Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **2021**, 6, 730-755. DOI: 10.1038/s41578-021-00340-w.

6. Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature.* **2023**, 624, 80-85. DOI: 10.1038/s41586-023-06735-9.

7. Feng, L. *et al.* Accelerated development of high-strength and high-conductivity Cu-Cr-Ti alloys based on data-driven design and experimental validation. *Mater. Des.* **2025**, 253, 113948. DOI: 10.1016/j.matdes.2025.113948.

8. Yin, J. *et al.* A novel neural network-based alloy design strategy: Gated recurrent unit machine learning modeling integrated with orthogonal experiment design and data augmentation. *Acta Mater.* **2023**, 243, 118420. DOI: 10.1016/j.actamat.2022.118420.

9. Zhang, H. T.; Fu, H. D.; Zhu, S. C.; Yong, W.; Xie, J. X. Machine learning assisted composition effective design for precipitation strengthened copper alloys. *Acta Mater.* **2021**, 215, 117118. DOI: 10.1016/j.actamat.2021.117118.

10. Zhang, H. T. *et al.* Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening. *Acta Mater.* **2020**, 200, 803-810. DOI: 10.1016/j.actamat.2020.09.068.

11 Li, L. J. *et al.* A machine learning strategy to achieve strength-conductivity-ductility synergy of high-performance Cu-Ni-Co-Si alloys via rolling and aging process. *J Mater. Sci. Technol.* **2026**, 267, 184-197.DOI: 10.1016/j.jmst.2025.12.040.

12. Sohail, Y. *et al.* Machine-learning design of ductile FeNiCoAlTa alloys with high strength. *Nature.* **2025**, 643, 119-124. DOI: 10.1038/s41586-025-09160-2.

13. Vazquez, G.; Chakravarty, S.; Gurrola, R.; Arróyave, R. A deep neural network regressor for phase constitution estimation in the high entropy alloy system Al-Co-Cr-Fe-Mn-Nb-Ni. *npj Comput. Mater.* **2023**, 9, 68. DOI: 10.1038/s41524-023-01021-8.

14. Li, H. C. *et al.* High-strength medium-entropy alloy designed by precipitation-strengthening mechanism via machine learning. *Mater. Sci. Eng. A.* **2023**, 882, 145443. DOI: 10.1016/j.msea.2023.145443.

15. Vela, B.; Khatamsaz, D.; Acemi, C.; Karaman, I.; Arróyave, R. Data-augmented modeling for yield strength of refractory high entropy alloys: A Bayesian approach. *Acta Mater.* **2023**, 261, 119351. DOI: 10.1016/j.actamat.2023.119351.

16. Wang, J.; Kwon, H.; Kim, H. S.; Lee, B. -J. A neural network model for high entropy alloy design. *npj Comput. Mater.* **2023**, 9, 60. DOI: 10.1038/s41524-023-01010-x.

17. Yin, J. *et al.* Interpretable predicting creep rupture life of superalloys: Enhanced by domain-specific knowledge. *Adv. Sci.* **2024**, 11, 2307982. DOI: 10.1002/advs.202307982.

18. Lian L. *et al.* Intelligent design of crack-resistant nickel-based superalloys for additive manufacturing by machine learning and multilayer filtering strategy. *Mater. Today Commun.* **2025**, 46, 112387. DOI: 10.1016/j.mtcomm.2025.112387.

19. Zhuang, X. L. *et al.* Alloying effects and effective alloy design of high-Cr CoNi-based superalloys via a high-throughput experiments and machine learning framework. *Acta Mater.* **2023**, 243, 118525. DOI: 10.1016/j.actamat.2022.118525.

20. Ma, Q. S. *et al.* Thermodynamic calculation and machine learning aided composition design of new nickel-based superalloys. *J. Mater. Res. Technol.* **2023**, 26, 4168-4178. DOI: 10.1016/j.jmrt.2023.08.139.
21. Yang, F. *et al.* Deep learning accelerates the development of Ni-based single crystal superalloys: A physical-constrained neural network for creep rupture life prediction. *Mater. Des.* **2023**, 232, 112174. DOI: 10.1016/j.matdes.2023.112174.
22. Xin, Y. *et al.* Building an effective deep learning model for mechanical properties prediction of steel. *Mater. Lett.* **2026**, 402, 139262. DOI: 10.1016/j.matlet.2025.139262.
23. Kannan, R.; Nandwana, P. Accelerated alloy discovery using synthetic data generation and data mining. *Scripta Mater.* **2023**, 228, 115335. DOI: 10.1016/j.scriptamat.2023.115335.
24. Wei, X. L.; van der Zwaag, S.; Jia, Z. X.; Wang, C. C.; Xu, W. On the use of transfer modeling to design new steels with excellent rotating bending fatigue resistance even in the case of very small calibration datasets. *Acta Mater.* **2022**, 235, 118103. DOI: 10.1016/j.actamat.2022.118103.
25. Ren, D.; Wang, C.; Wei, X.; Lai, Q.; Xu, W. Building a quantitative composition-microstructure-property relationship of dual-phase steels via multimodal data mining. *Acta Mater.* **2023**, 252, 118954. DOI: 10.1016/j.actamat.2023.118954.
26. Wu, J. *et al.* Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science.* **2024**, 386, 1256-1264. DOI: 10.1126/science.ads0901.
27. Xu, P. *et al.* Search for ABO_3 type ferroelectric perovskites with targeted multi-properties by machine learning strategies. *J. Chem. Inf. Model.* **2022**, 62, 5038-5049. DOI: 10.1021/acs.jcim.1c00566.
28. He, J. J. *et al.* Machine learning identified materials descriptors for ferroelectricity. *Acta Mater.* **2021**, 209, 116815. DOI: 10.1016/j.actamat.2021.116815.
29. Min, K.; Cho, E. Accelerated discovery of potential ferroelectric perovskite via active learning. *J. Mater. Chem. C.* **2020**, 8, 7866-7872. DOI: 10.1039/D0TC00985G.
30. Balachandran, P. V.; Kowalski, B.; Sehirlioglu, A.; Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **2018**, 9, 1668. DOI: 10.1038/s41467-018-03821-9.
31. Liu, Q. *et al.* Beyond designer's knowledge: Generating materials design hypotheses via a large language model. *Acta Mater.* **2025**, 297, 121307. DOI: 10.1016/j.actamat.2025.121307.
32. Tian, S. *et al.* Steel design based on a large language model. *Acta Mater.* **2025**, 285,

120663. DOI: 10.1016/j.actamat.2024.120663.
33. Wang, P. *et al.* Generalizable descriptors for automatic titanium alloys design by learning from texts via large language model. *Acta Mater.* **2025**, 296, 121275. DOI: 10.1016/j.actamat.2025.121275.
34. Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, 15, 1569. DOI: 10.1038/s41467-024-45914-8.
35. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, 577, 706-710. DOI: 10.1038/s41586-019-1923-7.
36. Tep, P.; Bernacki, M. High-fidelity grain growth modeling: Leveraging deep learning for fast computations. *Acta Mater.* **2025**, 301, 121486. DOI: 10.1016/j.actamat.2025.121486.
37. Yang, H. *et al.* Deep learning-based X-ray computed tomography image reconstruction and prediction of compression behavior of 3D printed lattice structures. *Addit. Manuf.* **2022**, 54, 102774. DOI: 10.1016/j.addma.2022.102774.
38. Lundberg, S. M.; Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 31, 4768-4777. DOI: 10.48550/arXiv.1705.07874.
39. Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. *arXiv.* **2016**. DOI: 10.48550/arXiv.1602.04938.
40. Alibagheri, E.; Ranjbar, A.; Khazaei, M.; Kühne, T. D.; Vaez Allaei, S. M. Remarkable optoelectronic characteristics of synthesizable square-octagon haeckelite structures: Machine learning materials discovery. *Adv. Funct. Mater.* **2024**, 34, 2470150. DOI: 10.1002/adfm.202402390.
41. Xu, H. Y. *et al.* Machine learning enabled the prediction of γ′-depleted depth during interdiffusion of bond-coated IN792 superalloy. *Surf. Coat Tech.* **2025**, 513, 132448. DOI: 10.1016/j.surfcoat.2025.132448.
42. Chen, W. *et al.* A map of single-phase high-entropy alloys. *Nat. Commun.* **2023**, 14, 2856. DOI: 10.1038/s41467-023-38423-7.
43. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, 2, 559-572. DOI: 10.1080/14786440109462720.
44. Srinivasan, S. *et al.* Mapping chemical selection pathways for designing multicomponent alloys: An informatics framework for materials design. *Sci. Rep.* **2015**, 5, 17960. DOI: 10.1038/srep17960.

45. Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science.* **2000**, 290, 2319-2323. DOI: 10.1126/science.290.5500.2319.

46. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv.* ***2020***. DOI: 10.48550/arXiv.1802.03426.

47. Zhao, S.; Li, J. S.; Wang J.; Lookman, T.; Yuan, R. H. Closed-loop inverse design of high entropy alloys using symbolic regression-oriented optimization. *Mater. Today.* **2025**, 88, 263-271. DOI: 10.1016/j.mattod.2025.06.033.

48. Xue, D. Z. *et al.* An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Mater.* **2017**, 125, 532-541. DOI: 10.1016/j.actamat.2016.12.009.

49. Pokorny, V. J.; Sponheim, S. R.; Rawls, E. Impact of reduced-dimensionality independent components analysis on event-related potential measurements. *Psychophysiology.* **2023**, 60, e14223. DOI: 10.1111/psyp.14223.

50. Miracle, D. B.; Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **2017**, 122, 448-511. DOI: 10.1016/j.actamat.2016.08.081.

51. Zhang, Y. *et al.* Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. *Acta Mater.* **2020**, 185, 528-539. DOI: 10.1016/j.actamat.2019.11.067.

52. Machaka, R.; Motsi, G. T.; Raganya, L. M.; Radingoana, P. M.; Chikosha, S. Machine learning-based prediction of phases in high-entropy alloys: A data article. *Data Brief.* **2021**, 38, 107346. DOI: 10.1016/j.dib.2021.107346.

53. Xue, D. Z. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **2016**, 7, 11241. DOI: 10.1038/ncomms11241.

54. Schölkopf, B.; Smola, A.; Müller, K. -R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **1998**, 10, 1299-1319. DOI: 10.1162/089976698300017467.

55. Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science.* **2000**, 290, 2323-2326. DOI: 10.1126/science.290.5500.2323.

56. Torgerson, W. S. Theory and methods of scaling. *J. Soc. Gen. Syst. Res.* **1958**, 4, 245-247. DOI: 10.1002/bs.3830040308.

57. Tao, Q. L., Xu, P. C., Li, M. J., Lu, W. C. Machine learning for perovskite materials design and discovery. *npj Comput. Mater.* **2021**, 7, 23. DOI: 10.1038/s41524-021-00495-8.

58. Zhang, Y., Zhou, Y. J., Lin, J. P., Chen, G. L., Liaw, P. K. Solid-Solution Phase Formation Rules for Multi-component Alloys. *Adv. Eng. Mater.*, **2008**, 10, 534-538. DOI: 10.1002/adem.200700240.

59. Zhang, Y. *et al.* Microstructures and properties of high-entropy alloys. *Prog. Mater. Sci.*, **2014**, 61, 1-93. DOI: 10.1016/j.pmatsci.2013.10.001.

60. Zarinejad, M.; Liu, Y. Dependence of transformation temperatures of NiTi-based shape-memory alloys on the number and concentration of valence electrons. *Adv. Funct. Mater.* **2008**, 18, 2789-2794. DOI: 10.1002/adfm.200701423.

61. Liu, Y. F.; Fu, X. Q.; Yu, Q.; Zhang, M. X.; Liu, J. Significant reduction of phase-transition hysteresis for magnetocaloric $(La_{1-x}Ce_x)_2Fe_{11}Si_2H_y$ alloys by microstructural manipulation. *Acta Mater.* **2021**, 207, 116687. DOI: 10.1016/j.actamat.2021.116687.

62. Zhou, Y. W. *et al.* Orchestrating phase transition in GeTe thermoelectrics: An investigation into the role of electronegativity. *Nano Energy.* **2024**, 127, 109723. DOI: 10.1016/j.nanoen.2024.109723.

63. Zurcher, R.; Muller, M.; Sachslehner, F.; Groger, V.; Zehetbauer, M. Dislocation resistivity in Cu: Dependence of the deviations from Matthiessen's rule on temperature, dislocation density and impurity content. *J. Phys-Condens. Mat.* **1995**, 7, 3515-3528. DOI: 10.1088/0953-8984/7/18/016.

64. Žnidarič, M. Modified Matthiessen's rule: More scattering leads to less resistance. *Phys. Rev. B* **2022**, 105, 45140. DOI: 10.1103/PhysRevB.105.045140.

65. Ho, C. Y. *et al.* Electrical resistivity of ten selected binary alloy systems. *J. Phys. Chem. Ref. Data.* **1983**, 12, 183-322. DOI: 10.1063/1.555684.

Supplementary Materials

Feature screening methods of high entropy alloy

A comprehensive dataset of high-entropy alloys was constructed by extracting composition and phase structure information from relevant literature[S1-S3]. To minimize the influence of heat treatment on phase formation, only data from alloys with as-cast microstructure obtained through arc melting were selected. The final dataset comprises 765 entries, including 4 ternary alloys, 106 quaternary alloys, 368 quinary alloys, 217 senary alloys and 70 alloys with seven or more components, encompassing 15 different elements. Alloy data with BCC, FCC, or FCC+BCC phase structures were labeled as "1", representing solid solution (SS) phases, while those with other phase structures were labeled as "-1", denoting non-solid solution (non-SS) phases. The dataset was subsequently divided into training and testing sets in an 80%:20% ratio for model training and predictive accuracy evaluation.

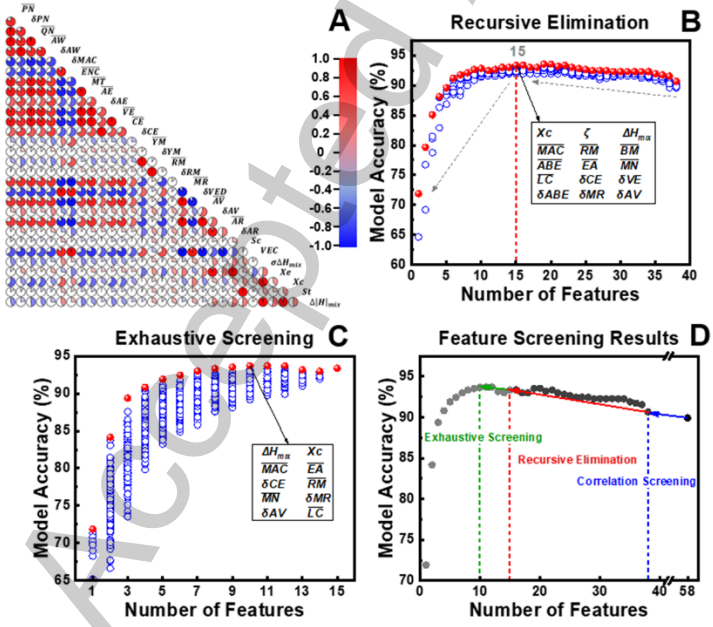
A comprehensive set of 22 elemental features was collected from the Pearson Handbook[S4] and literature[S5], including atomic number, electrochemical parameters, thermodynamic properties and size descriptors, as summarized in Supplementary Table 1. Based on the elemental composition and concentration data of each alloy sample, the mean value and standard deviation of these physicochemical features were calculated to represent the central tendency and variability of feature distribution across constituent elements. These statistical measures are defined as follows:

$$f_{mean} = \sum_{i=1}^{n} f_i \times \alpha_i \; / \; \sum_{i=1}^{n} \alpha_i \tag{S1}$$

$$f_{sd} = \sqrt{\sum_{i=1}^{n} (f_i - f_{mean})^2 \times \alpha_i \; / \; \sum_{i=1}^{n} \alpha_i} \tag{S2}$$

where f_i refers to the physicochemical feature of an element, i is the element index in the alloy, and α_i represents its atomic percentage content.

Furthermore, several empirical parameters relevant to high-entropy alloy phase classification were incorporated, including the ideal mixing entropy (ΔS_{mix}), total configurational entropy (ΔS_T), mixing enthalpy (ΔH_{mix}), standard deviation of mixing enthalpy ($\sigma\Delta H_{mix}$), absolute mixing enthalpy ($\Delta|H|_{mix}$), valence electron concentration (VEC), entropy-enthalpy ratio parameters (Ω, Ω_{mod}), entropy-enthalpy difference (ζ), atomic packing parameter (γ), geometric parameter (Λ), modulus mismatch (η), and potential energy quantification parameters (X_e, X_c)—totaling 14 empirical features. Subsequently, a three-step feature selection methodology[S6,S7] was employed to identify the most critical features influencing high-entropy alloy phase classification, as illustrated in Supplementary Figure 1.



Supplementary Figure 1. Feature selection results. (A) Pearson correlation screening; (B) Recursive elimination; (C) Exhaustive filtering; (D) The entire feature selection process.

Supplementary Table 1. Features used for high entropy alloys.

Feature	Definition	Feature	Definition
PN	Periodic number	VE	Vacancies enthalpy
QN	Quantum number	CE	Cohesive energy
AW	Atomic weight	YM	Young modulus
MAC	Mass attenuation coefficient	RM	Rigidity modulus
EP	Electronegativity (Pauling)	BM	Bulk modulus
ABE	Electronegativity absolute	MN	Mendeleev number
FIE	First ionization energy	MR	Metal radius
ENC	Effective nuclear charge	VED	Valence electron distance
EA	Electron affinity	AV	Atomic volume
MT	Melting temperature	LC	Lattice constants
AE	Atomization enthalpy	AR	Atomic radius

Evaluation indicators

The accuracy (Acc) is used to evaluate classification problems.

$$Acc=\frac{TP+TN}{TP+FN+FP+TN} \quad \text{(S3)}$$

Where, TP, TN, FP and FN represent True Positive, True Negative, False Positive and False Negative.

The average absolute percentage error (MAPE) and root mean square error (RMSE) are used to calculate the model prediction error:

$$MAPE=1/n\sum_{i=1}^{n}\left|y_{true\ i}-y_{pred\ i}\right| / y_{true\ i}\times 100\% \quad \text{(S4)}$$

$$RMSE=\sqrt{[\sum(y_{true\ i}-y_{pred\ i})^2/n]} \quad \text{(S5)}$$

Where, $y_{pred\ i}$ and $y_{true\ i}$ represent the predicted value and actual value of alloy i respectively, and n represents the number of alloys in the subset.

Experimental procedures

Refractory high-entropy alloys were synthesized by vacuum arc melting in a water-cooled copper crucible using high-purity (＞99.99%) elements: V, Nb, Ti,

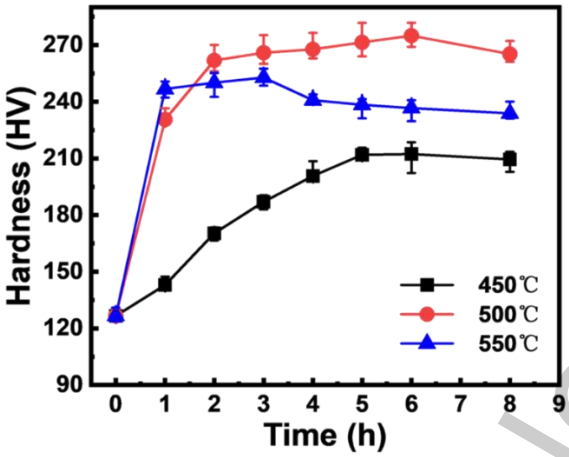
W, Zr, Mo, Hf, and Ta. To achieve composition homogeneity, the melting process was repeated eight times. Phase characterization was conducted by X-ray diffraction (XRD) with a 2θ range of 20° to100° and a scanning speed of 4° min^{-1}.

The design space for copper alloys is presented in Supplementary Table 2. Guided by prior knowledge[S6], the content of Ni, Co, and Si in Cu-Ni-Si and Cu-Ni-Co-Si alloys was constrained to maintain the (Ni+Co)/Si ratio (in wt%) between 4 and 5, thereby effectively narrowing down the alloy composition candidate space.

Supplementary Table 2. Alloying elements content ranges.

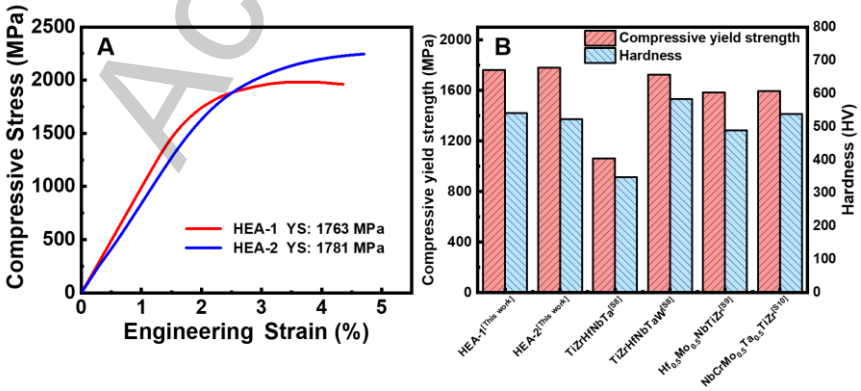
Elements	Cu	Ni	Si	Co	Mg	P	Zr	Zn
Content (wt.%)	Bal.	0-9	0-1.5	0-1.5	0-0.3	0-0.1	0-0.3	0-0.3

The copper alloys were prepared using medium-frequency vacuum induction melting. The resulting ingots were homogenized at 900 °C for 4 h, hot-rolled at 850 °C with a 90% reduction, followed by solution treatment at 975 °C for 2 h and subsequent aging. Aging treatments were conducted at temperatures of 450 °C, 500 °C and 550 °C, each for durations of 1, 2, 3, 4, 5, 6 and 8 h. Vickers hardness was measured using an HXD-1000T hardness tester under a load of 100 gf with a dwell time of 15 s; each sample was tested five times and the average value was recorded as the hardness. Electrical resistivity at peak hardness was determined at room temperature with an Applent AT-510Pro DC resistance tester. The variation of alloy hardness with aging time is shown in Supplementary Figure 2.



Supplementary Figure 2. Variation curve of alloy hardness and aging time.

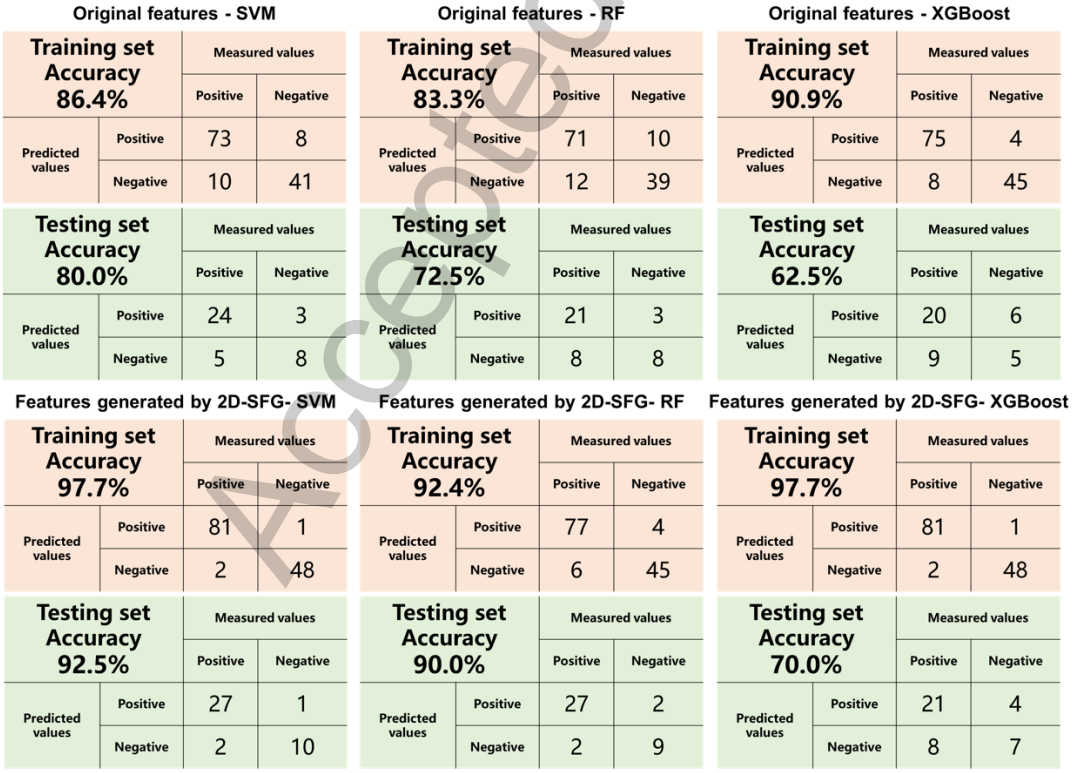
For refractory high-entropy alloys, prepare cylindrical specimens of Φ4x5 dimensions. The tested microhardness was assessed on polished cross-sections under a 0.5N force for 30 s. The Vickers hardness experiments were repeated five times for each examined sample to get the average values. Conduct compression tests on the prepared specimens using an MTS-SANS CMT5000 series microcomputer-controlled electronic universal testing machine. The compression temperature is set at 25°C, and the compression rate is maintained at 10^{-3}·s^{-1} until fracture occurs, in order to test the room temperature yield strength of the alloy. The corresponding stress-strain curves and room temperature mechanical properties are provided in Supplementary Figure 3.



Supplementary Figure 3. (A) Stress-strain curves of RHEAs; (B). Room temperature mechanical properties of RHEAs.

Algorithm-Agnostic Nature of the 2D-SFG Framework

To verify the algorithm-agnostic nature of the 2D-SFG method, we constructed machine learning models based on different algorithms using both 'original features' and '2D-SFG-generated features' on the same datasets, and compared their predictive capabilities. The training and prediction results for all models are presented in Supplementary Figures 4 to 8, while the cross-validation performance metrics are summarized in Supplementary Table 3. Across both classification and regression tasks, it is evident that regardless of whether SVM, RF, or XGBoost was employed, the models utilizing 2D-SFG-generated physical features achieved significantly improved predictive performance compared to those built on original features. These findings demonstrate that the 2D-SFG method is indeed algorithm-agnostic; the features it generates encode richer physical information, thereby universally enhancing the performance of various mainstream machine learning algorithms.



Supplementary Figure 4. Confusion matrix of different models in perovskite classification task.

Original features - SVM

Training set Accuracy 88.6%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	286	27
	Negative	36	206

Testing set Accuracy 86.6%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	99	13
	Negative	15	83

Original features - RF

Training set Accuracy 85.9%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	277	33
	Negative	45	200

Testing set Accuracy 78.1%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	88	20
	Negative	26	76

Original features - XGBoost

Training set Accuracy 91.7%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	296	20
	Negative	26	213

Testing set Accuracy 73.3%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	84	26
	Negative	30	70

Features generated by 2D-SFG- SVM

Training set Accuracy 91.9%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	297	20
	Negative	25	213

Testing set Accuracy 91.4%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	104	8
	Negative	10	88

Features generated by 2D-SFG- RF

Training set Accuracy 89.4%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	289	26
	Negative	33	207

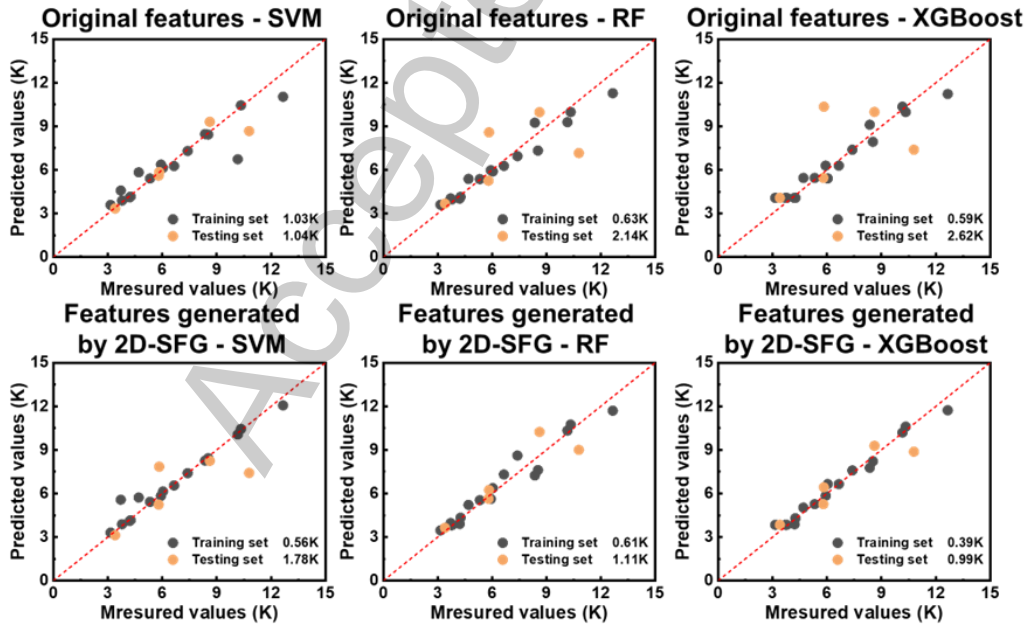
Testing set Accuracy 82.9%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	94	16
	Negative	20	80

Features generated by 2D-SFG- XGBoost

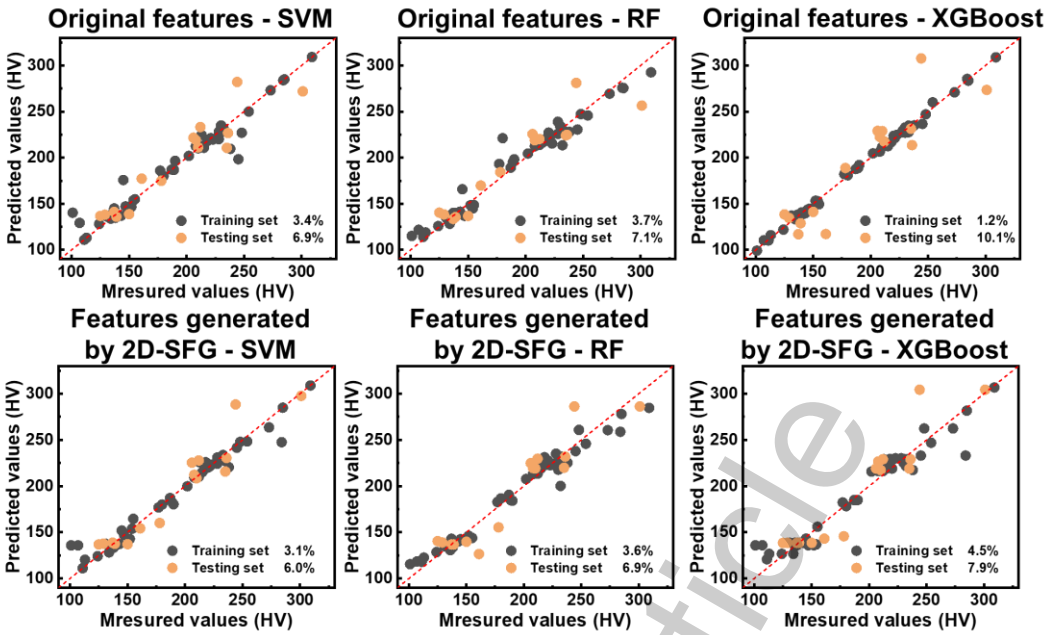
Training set Accuracy 94.8%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	303	10
	Negative	19	223

Testing set Accuracy 71.9%		Measured values: Positive	Measured values: Negative
Predicted values	Positive	77	22
	Negative	37	74

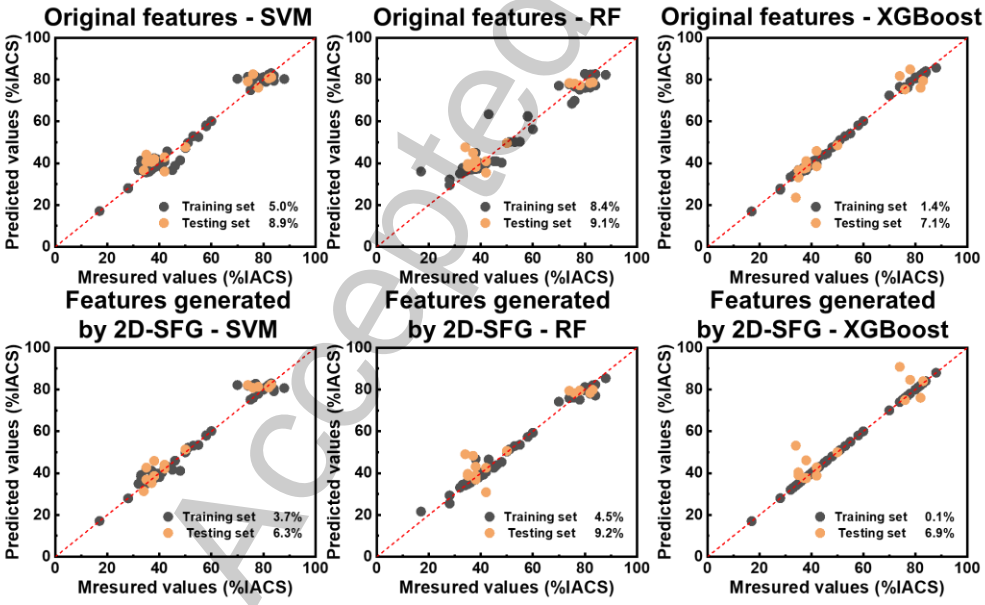
Supplementary Figure 5. Confusion matrix of different models in high entropy alloy classification task.



Supplementary Figure 6. Predictive performance of different models in the shape memory alloy regression task.



Supplementary Figure 7. Predictive performance of different models in the copper alloy hardness regression task.



Supplementary Figure 8. Predictive performance of different models in the copper alloy hardness regression task.

Supplementary Table 3. The predictive performance of different algorithms and features.

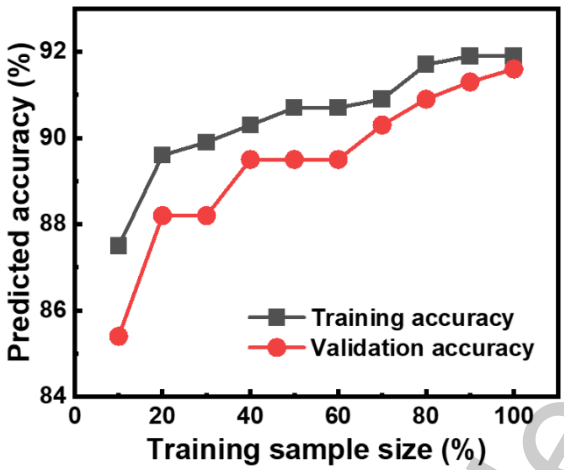
Methods	Perovskites (ACC)	High-entropy alloys (ACC)	Thermal hysteresis (RMSE)	Hardness (MAPE)	Electrical conductivity (MAPE)
OF-SVR	85.6%	84.5%	2.9K	5.9%	9.7%
SFG-SVR	94.2%	88.4%	1.1K	3.7%	6.2%
OF-RF	83.2%	80.5%	2.7K	9.6%	12.4%
SFG-RF	90.7%	83.9%	2.3K	5.3%	7.5%
OF-XGBoost	88.1%	81.7%	1.7K	4.4%	11.2%
SFG-XGBoost	90.1%	83.3%	1.3K	3.1%	6.9%

Note: OF - original features, SFG - features generated by 2D-SFG. The data in the table represents the 10-fold cross-validation accuracy/error of the model.

Overfitting risk and model stability assessment

To comprehensively evaluate the risk of overfitting and model stability, we constructed models using varying sizes of training sets and generated learning curves for the representative task of high-entropy alloy phase classification. As shown in Supplementary Figure 9, both training and validation accuracies exhibit an upward trend and gradually converge as the training sample size increases. Crucially, the gap between the two metrics remains consistently small throughout, indicating that the model is well-fitted and does not suffer from significant overfitting.

Furthermore, we conducted 100 repetitions of stratified random splitting. The results demonstrate stable predictive performance, with a training accuracy of 92.1% ± 0.7% and a test accuracy of 91.3% ± 0.9%. These findings confirm that the model's performance is insensitive to the randomness of data partitioning, exhibiting high robustness.



Supplementary Figure 9. Learning curves for phase classification tasks.

Considering the high computational cost associated with symbolic regression, we employed 5 repetitions of stratified sampling. We analyzed the symbolic features generated across these five independent runs, with results detailed in Supplementary Table 4. Although the exact mathematical expressions were not identical in every run, key features and specific combination forms (e.g., δMR×δAV and δAV/δCE) recurred consistently in the vast majority of trials. This demonstrates that the physical laws uncovered by the 2D-SFG method possess intrinsic consistency and are not artifacts of random noise.

Supplementary Table 4. Results of 5 independent symbol feature generations.

No.	symbol feature	The top 5 features that appear most frequently in 100 iterations
1	$X_{C3} = \delta MR \times \delta AV \times \left(2 \times \overline{MAC} \times \frac{\overline{EA}}{\overline{LC}} + \overline{RM} + \delta AV\right)$ $X_{C4} = \overline{MN} - Xc + \frac{\delta AV}{\delta CE} + \frac{\overline{MN}}{\overline{EA}}$	δMR, $\overline{RM}$, δCE, $\overline{MN}$, δAV
2	$X_{C3} = \delta MR \times \delta AV + \overline{RM}$ $X_{C4} = \delta AV + \frac{\delta AV}{\delta CE} + \frac{\overline{MAC}}{\overline{LC} \times Xc}$	δMR, $\overline{RM}$, δCE, $\overline{MAC}$, δAV

3	$X_{C3} = \delta MR \times \delta AV + \frac{\overline{MN}}{\overline{RM}}$ $X_{C4} = \frac{\overline{EA}}{\overline{LC} \times \overline{MN}} - \frac{\overline{EA}}{\overline{\delta CE}}$	$\delta MR, \overline{RM}, \delta CE, \overline{MN}, \overline{EA}$
4	$X_{C3} = \delta MR \times (\overline{MAC} \times \overline{EA} - \overline{LC})$ $X_{C4} = \overline{MN} \times \frac{\delta AV}{\delta CE}$	$\delta MR, \overline{RM}, \delta CE, \overline{MN}, \delta AV$
5	$X_{C3} = \delta MR \times \delta AV - \frac{\delta AV}{\overline{EA}} + \overline{RM}$	$\delta MR, \overline{RM}, \delta CE, \overline{MN}, \overline{EA}$

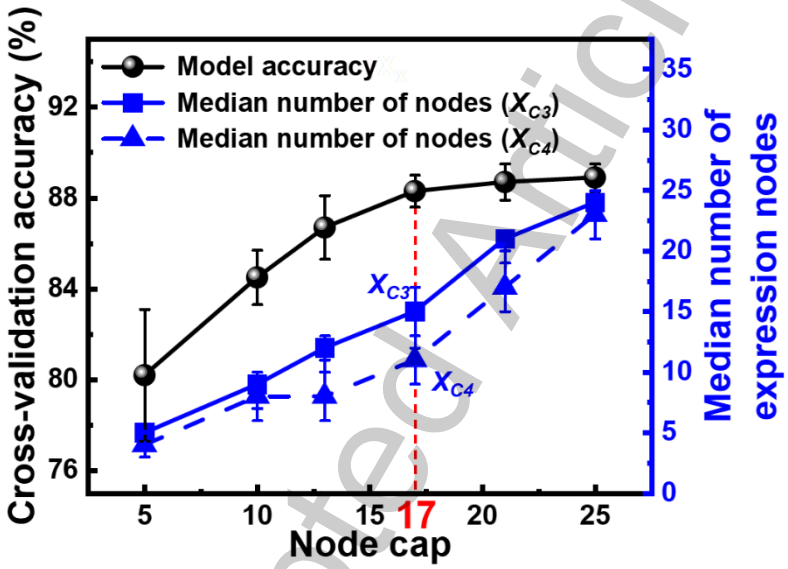
Sensitivity Analysis on Expression Complexity

To determine the optimal constraint on the number of nodes in symbolic regression trees, a sensitivity analysis using the high-entropy alloy phase classification task as a representative case was conducted. The maximum node limit was varied across six values: {5, 10, 13, 17, 21, 25}. To mitigate the impact of the stochastic inherent in genetic algorithms, five independent runs with distinct random seeds for each setting was performed and both the cross-validation accuracy and the median number of nodes in the final generated formula trees were recorded.

As illustrated in Supplementary Figure 10, the model was overly constrained when maximum node cap was set to 5, yielding a low mean accuracy of 80.2% ± 2.9%. As the node cap increased to 17, the cross-validation accuracy improved significantly, reaching 88.3% ± 0.7%. Beyond this point, further increases in the cap resulted in a plateau in the accuracy curve, indicating that additional expressive capacity did not translate into further gains in predictive power. The error bars in the figure represent the standard deviation of the prediction results across the five independent runs. The small fluctuations in accuracy across different random seeds confirm the stability of the performance improvement.

Furthermore, a higher allowed node count led to an increase in the median number of nodes in the generated formulas. Notably, when the cap exceeded 17, formula complexity rose sharply, resulting in significantly longer expressions

that compromised model interpretability. Similarly, the error bars for the median node count (representing the fluctuation range across the five runs) were short. This demonstrates that, despite the stochasticity of the genetic algorithm, the complexity of the formulas generated by our 2D symbolic feature method remained highly consistent across multiple runs, avoiding unstable variations where formula lengths fluctuate wildly between runs. In summary, 17 is identified as the optimal trade-off point, balancing predictive performance with formula complexity.



Supplementary Figure 10. Sensitivity analysis of the maximum node cap. Error bars indicate the standard deviation / fluctuation range across these runs, demonstrating the stability of both predictive performance and model complexity.

REFERENCES

S1. Miracle, D. B.; Senkov, O. N. A critical review of high entropy alloys and related concepts. Acta Mater. 2017, 122, 448-511. DOI: 10.1016/j.actamat.2016.08.081.

S2. Zhang, Y. et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. Acta Mater. 2020, 185, 528-539. DOI: 10.1016/j.actamat.2019.11.067.

S3. Machaka, R.; Motsi, G. T.; Raganya, L. M.; Radingoana, P. M.; Chikosha, S. Machine learning-based prediction of phases in high-entropy alloys: A data article. Data Brief. 2021, 38, 107346. DOI: 10.1016/j.dib.2021.107346.

S4. Pearson, K. Note on regression and inheritance in the case of two parents. Proc. R. Soc. London. 1895, 58, 240-242. DOI: 10.1098/rspl.1895.0041.
S5. Villars, P. et al. Binary, ternary and quaternary compound former/nonformer prediction via mendeleev number. J. Alloy Compd. 2001, 317, 26-38. DOI: 10.1016/s0925-8388(00)01410-9.
S6. Zhang, H. T.; Fu, H. D.; Zhu, S. C.; Yong, W.; Xie, J. X. Machine learning assisted composition effective design for precipitation strengthened copper alloys. Acta Mater. 2021, 215, 117118. DOI: 10.1016/j.actamat.2021.117118.
S7. Zhang, H. T. et al. Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening. Acta Mater. 2020, 200, 803-810. DOI: 10.1016/j.actamat.2020.09.068.
S8. Huang, W. J., Wang, X. J., Qiao, J. W., Wu, Y. C. Microstructures and mechanical properties of TiZrHfNbTaWx refractory high entropy alloys. J Alloy Compd. 2022, 914, 165187. DOI: 10.1016/j.jallcom.2022.165187.
S9. Guo, N. N. et al. Microstructure and mechanical properties of in-situ MC-carbide particulates-reinforced refractory high-entropy Mo0.5NbHf0.5ZrTi matrix alloy composite. Intermetallics, 2016, 69, 74-77. DOI: 10.1016/j.intermet.2015.09.011.
S10. Senkov, O. N. and Woodward, C. F. Microstructure and properties of a refractory NbCrMo0.5Ta0.5TiZr alloy. *Mater. Sci. Eng. A.* **2011**, 529, 311-320. DOI: 10.1016/j.msea.2011.09.033.