Supplementary Materials

Intelligent visualization-driven materials design via two-dimensional symbolic feature generation

Wei Yong[1,#], Hongtao Zhang[1,2,3,4,#,*], 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]Key Laboratory for Advanced Materials Processing (MOE), School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China.
[5]Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, Liaoning, China.
[#]Authors contributed equally to this work.

[*]**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

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:

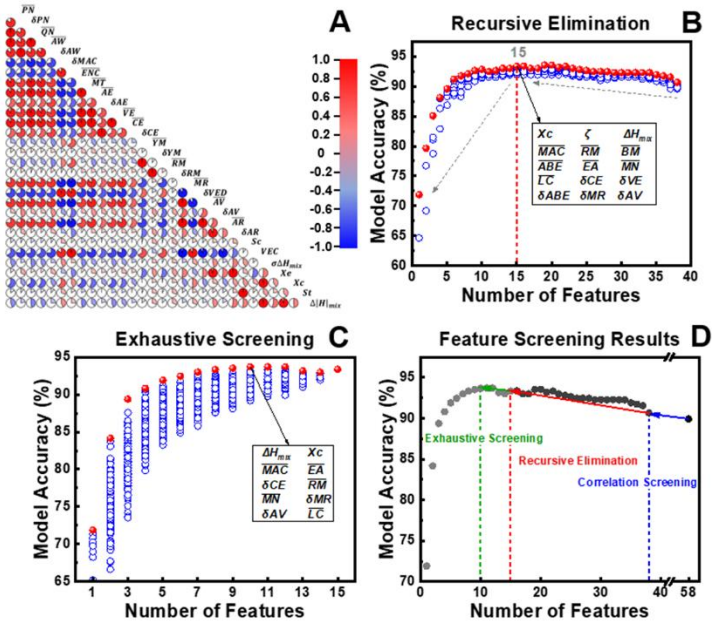
$$f_{mean} = \sum_{i=1}^{n} f_i \times \alpha_i \; / \; \sum_{i=1}^{n} \alpha_i \tag{S1}$$

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

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.

$$\mathbf{Acc}=\frac{\mathbf{TP+TN}}{\mathbf{TP+FN+FP+TN}} \tag{S3}$$

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:

$$\mathbf{MAPE}=1/n\sum_{i=1}^{n}\left|y_{true\ i}-y_{pred\ i}\right| / y_{true\ i}\times 100\% \tag{S4}$$

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

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

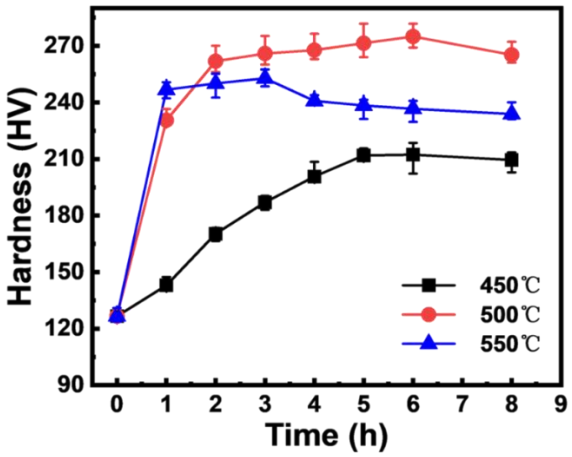
Refractory high-entropy alloys were synthesized by vacuum arc melting in a water-cooled copper crucible using high-purity (＞99.99%) elements: V, Nb, Ti, W, Zr, Mo, Hf, and Ta. To achieve composition homogeneity, the melting process was repeated eight times. Phase characterization was conducted by X-ray 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

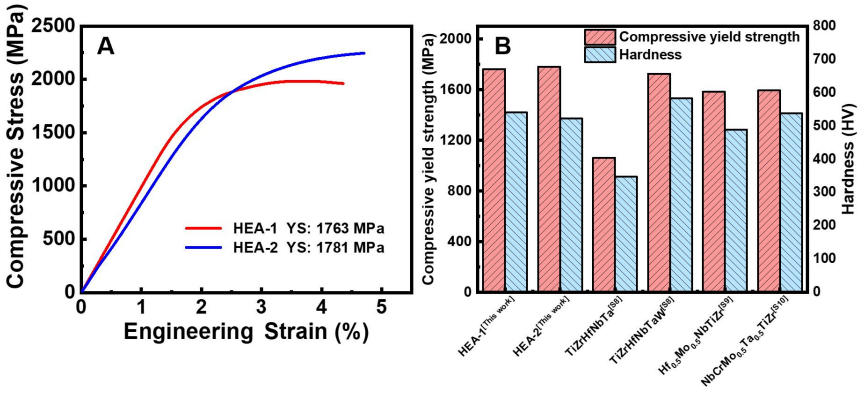
The copper alloys were prepared using medium-frequency vacuum induction melting. The resulting ingots were homogenized at 900 °C for 4 h, hot-rolled at 850 °C with a 90% reduction, followed by solution treatment at 975 °C for 2 h and subsequent aging. Aging treatments were conducted at temperatures of 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

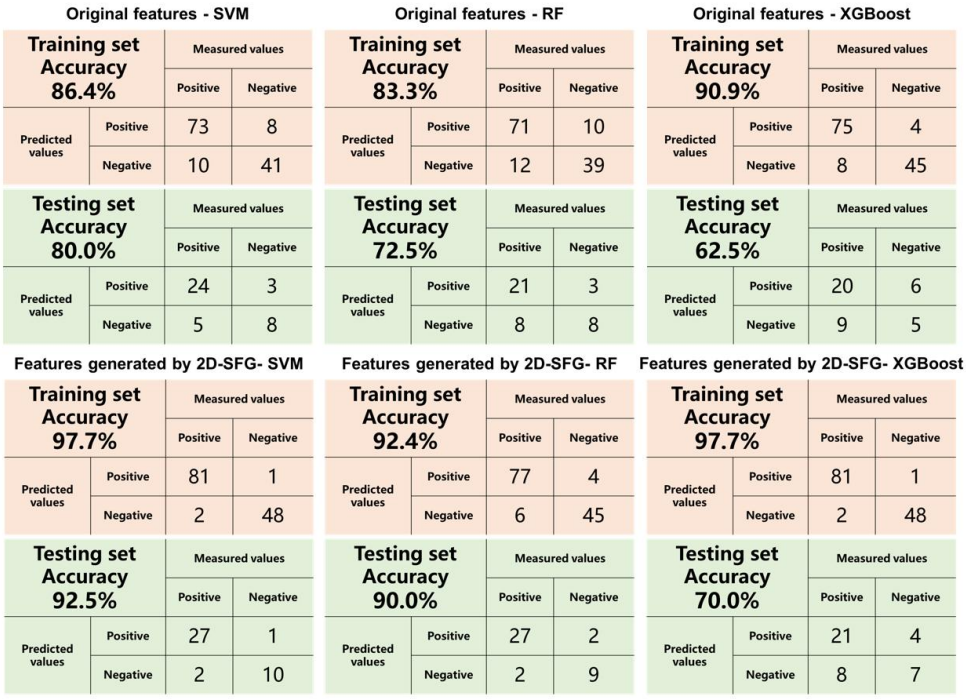
electronic universal testing machine. The compression temperature is set at 25°C, and the compression rate is maintained at $10^{-3} \cdot s^{-1}$ until fracture occurs, in order to test the room temperature yield strength of the alloy. The corresponding stress-strain curves and room temperature mechanical properties are provided in Supplementary Figure 3.



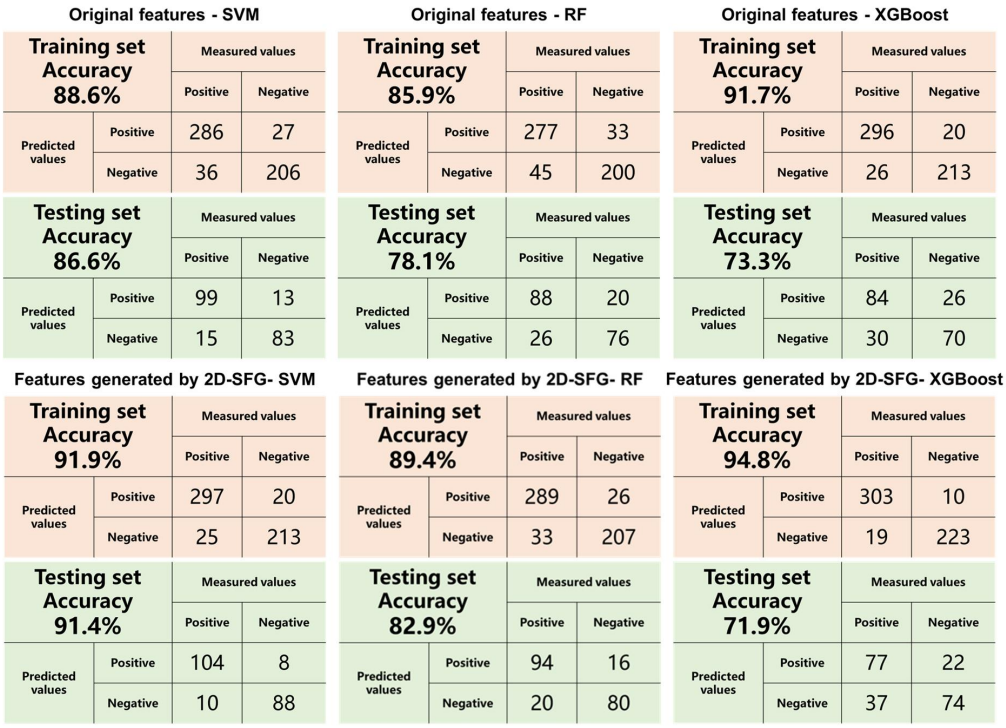
Supplementary Figure 3. (A) Stress-strain curves of RHEAs; (B). Room temperature mechanical properties of RHEAs.

Algorithm-Agnostic Nature of the 2D-SFG Framework

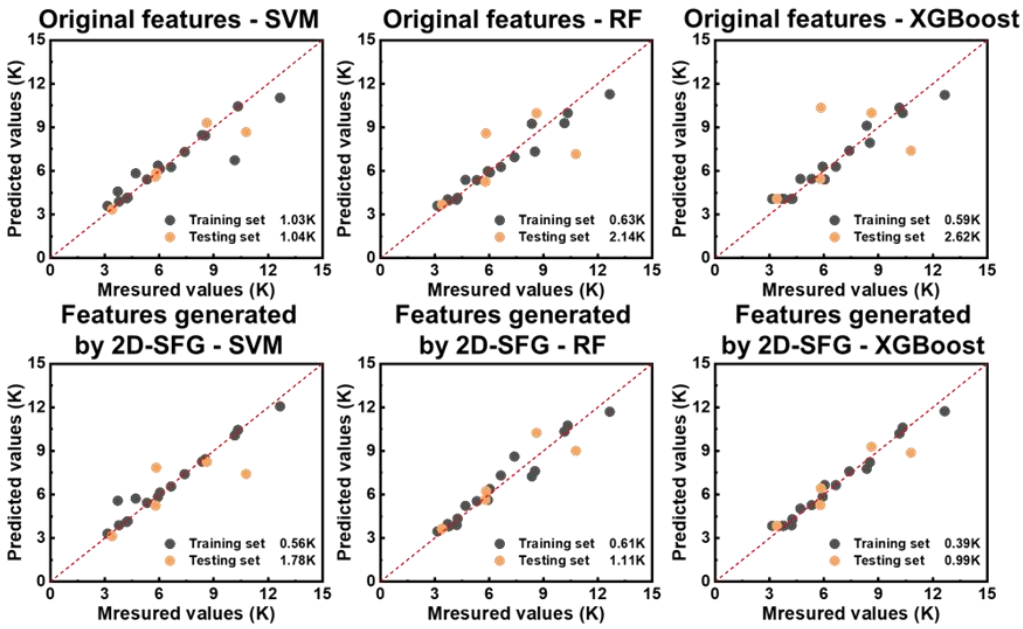
To verify the algorithm-agnostic nature of the 2D-SFG method, we constructed machine learning models based on different algorithms using both 'original features' and '2D-SFG-generated features' on the same datasets, and compared their predictive capabilities. The training and prediction results for all 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.



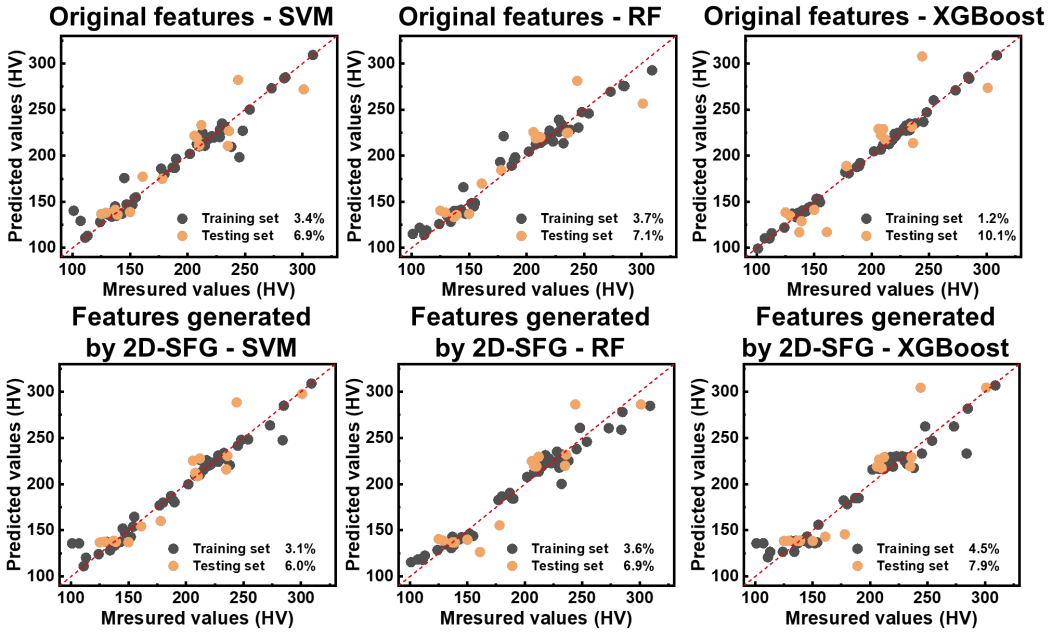
Supplementary Figure 4. Confusion matrix of different models in perovskite classification task.



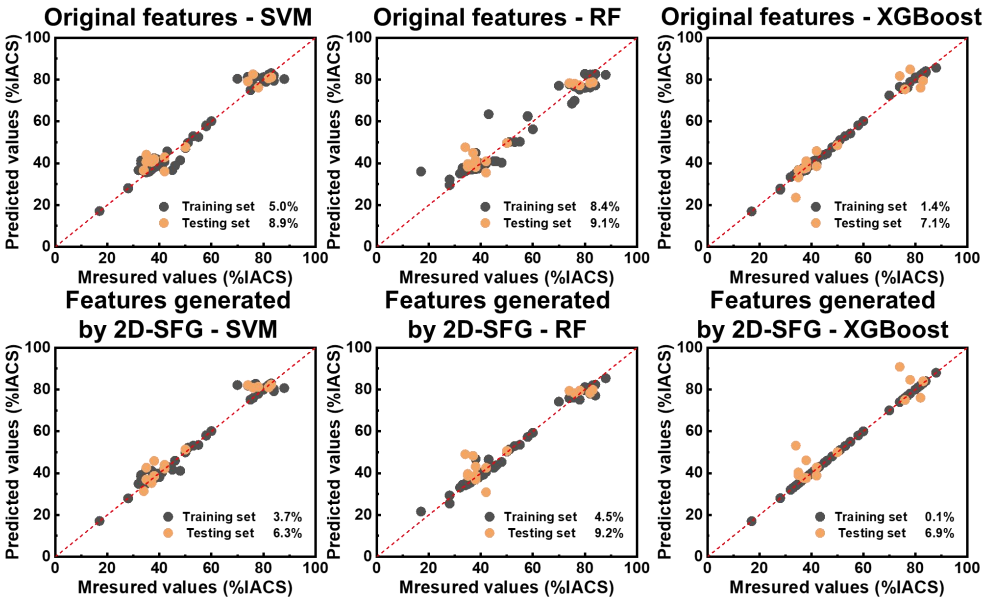
Supplementary Figure 5. Confusion matrix of different models in high entropy alloy classification task.



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.

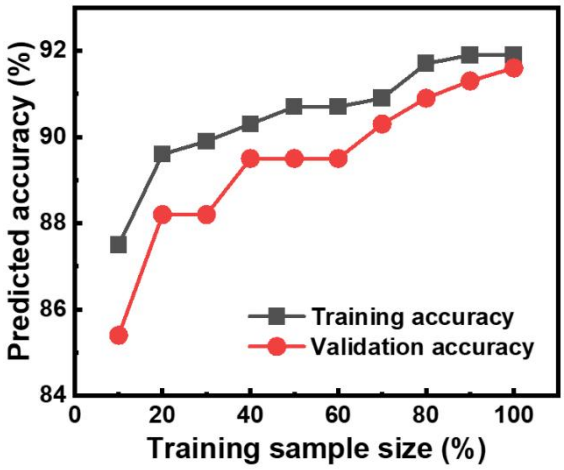
Methods	Perovskites (ACC)	High-entropy alloys (ACC)	Thermal hysteresis (RMSE)	Hardness (MAPE)	Electrical conductivity (MAPE)
OF-SVR	85.6%	84.5%	2.9K	5.9%	9.7%
SFG-SVR	94.2%	88.4%	1.1K	3.7%	6.2%
OF-RF	83.2%	80.5%	2.7K	9.6%	12.4%
SFG-RF	90.7%	83.9%	2.3K	5.3%	7.5%
OF-XGBoost	88.1%	81.7%	1.7K	4.4%	11.2%
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., $\delta MR \times \delta AV$ and $\delta AV/\delta CE$) 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	$X_{C3}=\delta MR\times\delta AV\times\left(2\times\overline{MAC}\times\frac{\overline{EA}}{\overline{LC}}+\overline{RM}+\delta AV\right)$ $X_{C4}=\overline{MN}\text{-}Xc+\frac{\delta AV}{\delta CE}+\frac{\overline{MN}}{\overline{EA}}$	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	$X_{C3}=\delta MR\times\delta AV+\frac{\overline{MN}}{\overline{RM}}$ $X_{C4}=\frac{\overline{EA}}{\overline{LC}\times\overline{MN}}\text{-}\frac{\overline{EA}}{\overline{\delta CE}}$	δMR, $\overline{RM}$, δCE, $\overline{MN}$,$\overline{EA}$
4	$X_{C3}=\delta MR\times(\overline{MAC}\times\overline{EA}-\overline{LC})$ $X_{C4}=\overline{MN}\times\frac{\delta AV}{\delta CE}$	δMR, $\overline{RM}$, δCE, $\overline{MN}$, δAV
5	$X_{C3}=\delta MR\times\delta AV\text{-}\frac{\delta AV}{\overline{EA}}+\overline{RM}$	δMR, $\overline{RM}$, δCE, $\overline{MN}$, $\overline{EA}$

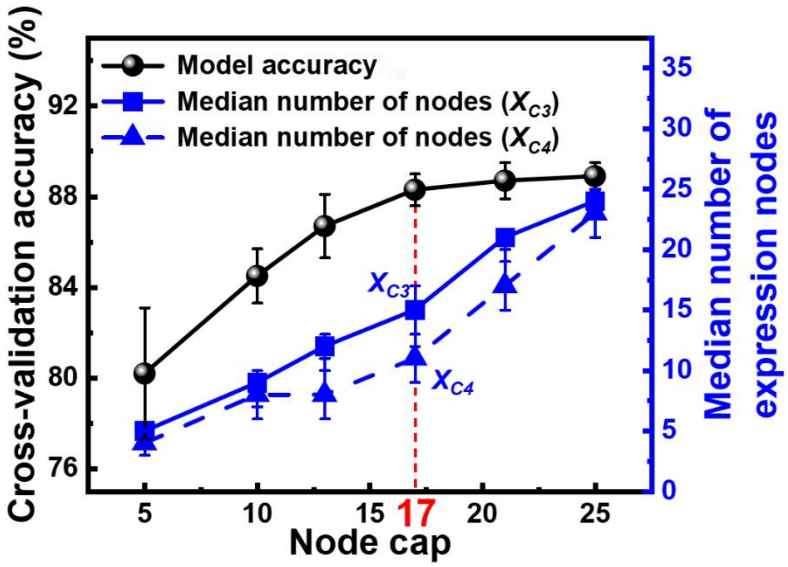
Sensitivity Analysis on Expression Complexity

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.