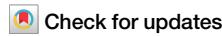




# Corrosion resistance prediction of high-entropy alloys: framework and knowledge graph-driven method integrating composition, processing, and crystal structure



Guangxuan Song<sup>1</sup>✉, Dongmei Fu<sup>1</sup>✉, Yongjie Lin<sup>2</sup>✉, Lingwei Ma<sup>2</sup>✉ & Dawei Zhang<sup>2</sup>✉

The prediction of corrosion resistance in High-entropy alloys (HEAs) faces challenges due to previous machine learning methods not fully capturing the interdependencies between composition, processing, and crystal structure. This study proposes the Composition and Processing-Driven Two-Stage Corrosion Prediction Framework with Structural Prediction (CPSP Framework), which first predicts crystal structure and then combines composition and processing data for corrosion current prediction. A deep learning model, Mat-NRKG, is developed based on the CPSP framework, efficiently integrating composition, processing, and crystal structure data through a knowledge graph and graph convolutional network. Evaluations using the HEA-CRD dataset show that the CPSP Framework outperforms the Composition-Only Prediction Framework (CP Framework) and the Composition and Processing-Based Prediction Framework (CPP Framework). The Mat-NRKG model demonstrates the best performance on the HEA-CRD dataset. Its generalization capability is validated through experiments on five laboratory-synthesized HEAs, highlighting the effectiveness of incorporating prior knowledge into model design for performance prediction.

High-entropy alloys (HEAs)<sup>1,2</sup> represent an emerging class of metallic materials, characterized by their disordered chemical environments and high mixed entropy. These characteristics contribute to their exceptional mechanical properties<sup>3,4</sup>, oxidation resistance<sup>5,6</sup>, and corrosion resistance<sup>7–11</sup>. Corrosion resistance plays a crucial role in the performance and lifespan of HEAs in harsh environments<sup>12</sup>. While chemical composition is a key factor, corrosion resistance is, in many cases, also influenced by microstructure and processing techniques<sup>13–15</sup>. In particular, the crystal structure influences elemental distribution and phase stability, which in turn affects localized corrosion behavior<sup>16</sup>.

Due to the vast compositional space<sup>17,18</sup>, developing HEAs using traditional experimental approaches often face high costs and time constraints<sup>19,20</sup>. Although computational simulations<sup>21–23</sup> can provide theoretical guidance for materials design, they typically require extensive computational resources<sup>24</sup> and are limited to small scales, making it difficult

to fully capture the effects of composition, structure, processing conditions, and complex environmental effects on the corrosion resistance of HEAs<sup>25</sup>.

Machine learning (ML) techniques have capability of uncovering underlying patterns from experimental data and predicting material properties<sup>26–28</sup>, offering an effective approach to HEA design and optimization<sup>29–32</sup>. Although some ML-based studies have been conducted on HEA performance prediction, limitations remain in current methods. Some studies<sup>13,33,34</sup> focus solely on the chemical composition's effect on corrosion resistance, neglecting the influence of non-compositional factors such as processing techniques and crystal structure. The study<sup>35</sup> has integrated composition, processing, and structure into machine learning models. This integration of prior knowledge enhances the interpretability of the predictive models<sup>36,37</sup>. However, using structural information as an explicit input often limits the model's engineering applicability, since obtaining such data typically requires experimental preparation and

<sup>1</sup>Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. <sup>2</sup>National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing, China.

✉ e-mail: [suguangxuan@xs.ustb.edu.cn](mailto:suguangxuan@xs.ustb.edu.cn); [fmd\\_ustb@ustb.edu.cn](mailto:fmd_ustb@ustb.edu.cn); [b20200299@xs.ustb.edu.cn](mailto:b20200299@xs.ustb.edu.cn); [mlw1215@ustb.edu.cn](mailto:mlw1215@ustb.edu.cn); [dzhang@ustb.edu.cn](mailto:dzhang@ustb.edu.cn)

characterization<sup>38,39</sup>, theoretical modeling<sup>40–42</sup>, or simulation<sup>43,44</sup>. Therefore, building models that can simultaneously capture the complex relationships among composition, processing, and structure, while maintaining high interpretability, predictive performance, and practical usability, remains a significant challenge.

To address this challenge, an innovative the Composition and Processing-Driven Two-Stage Corrosion Prediction Framework with Structural Prediction (CPSP Framework) is proposed for the first time, which hierarchically models the composition – processing – structure – performance relationships. By first predicting the crystal structure and then integrating it with composition and processing data, CPSP eliminates the need for experimentally obtained structural input during the inference stage, thereby improving its engineering applicability. Furthermore, the framework is compatible with various ML models capable of both classification and regression, ensuring broad adaptability. Building upon the CPSP framework, a specialized deep learning model, Mat-NRKG, is proposed, leveraging knowledge graph techniques<sup>45</sup> for enhanced predictive capabilities. The knowledge graph organizes and models unstructured process-related data from the literature in a flexible graph structure<sup>46</sup>, which facilitates the capture of capturing complex relationships among composition, processing, and structure-performance correlations. The Mat-NRKG model leverages the TransE algorithm<sup>47</sup> for knowledge graph completion to predict crystal structure, while integrating compositional and processing information through a Graph Convolutional Network (GCN)<sup>48</sup> augmented with a Deep Taylor Block (DTB)<sup>49</sup> module. Ultimately, this end-to-end

approach predicts corrosion current. In this study, we focus on improving precision, referring to the consistency of our model's predictions, rather than accuracy, which is influenced by the quality of the input data. By leveraging a knowledge graph to encode the interactions among composition, processing, and predicted structure, Mat-NRKG enables more transparent reasoning and provides a certain level of interpretability, while also improving prediction precision. Compared to the original NRKG model, which requires crystal structure as an input, Mat-NRKG incorporates a structure prediction step based on composition and processing information, making it more applicable in practical settings where structural data may be unavailable, and thereby potentially increasing its engineering utility. The Composition-Only Prediction Framework (CP Framework) and the Composition and Processing-Based Prediction Framework (CPP Framework) are constructed as baselines to evaluate the CPSP Framework and Mat-NRKG model. The HEA-CRD dataset<sup>35</sup> is used to compare the corrosion resistance prediction performance of these models, where CPSP demonstrates consistent improvements over CP and CPP, while Mat-NRKG further enhances performance, reducing MSE by at least 25%. Additionally, five HEAs were synthesized to validate generalization, demonstrating robustness of the CPSP framework and Mat-NRKG method.

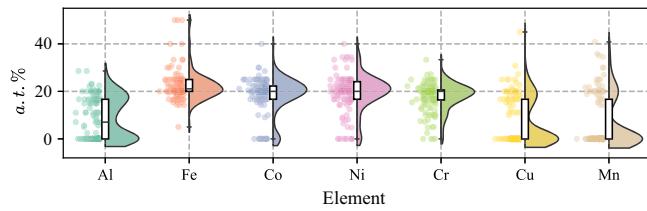
## Results and discussion

### HEA corrosion resistance dataset and statistical analysis

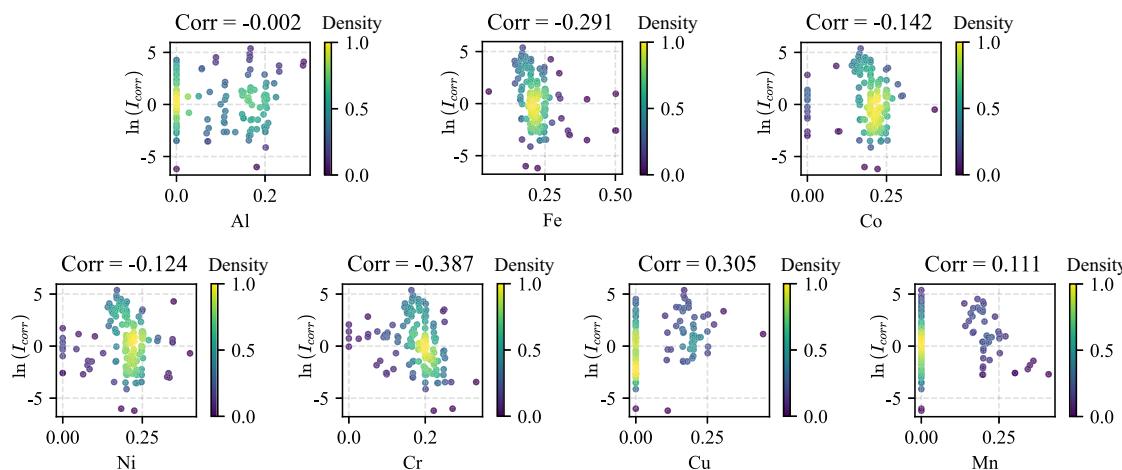
The existing literature provides extensive data on the corrosion resistance of HEAs, but typically reports only the corrosion current for a single alloy or a specific alloy category. The dataset used in this study is the HEA Corrosion Resistance Dataset (HEA-CRD)<sup>35</sup>, which was curated by combining large language models with manual inspection to select 151 corrosion resistance records extracted from the literature. These data encompass the composition, processing techniques, and crystal structures of HEAs within the Al-Co-Cr-Fe-Cu-Ni-Mn system. Corrosion current densities were measured at 25 °C (or room temperature) in a 3.5 wt% NaCl solution using polarization experiments.

Figure 1 shows the distribution of alloy compositions for Al, Co, Cr, Fe, Cu, Ni, and Mn. Alloys containing Fe, Co, Ni, Cr, Cu, or Mn typically have atomic percentages of these elements ranging from 10 to 30%, while the atomic percentage of Al is generally below 20%. Additionally, Al, Cu, and Mn often exhibit distinct bimodal distributions in many HEAs.

Figure 2 shows the Spearman correlation analysis<sup>50</sup> between elements and material properties. It indicates that Cr and Cu exhibit correlation coefficients (absolute value) with corrosion current ( $\ln(I_{\text{corr}})$ ) greater than 0.3, suggesting moderate correlation with  $\ln(I_{\text{corr}})$ <sup>51</sup>. In contrast, other

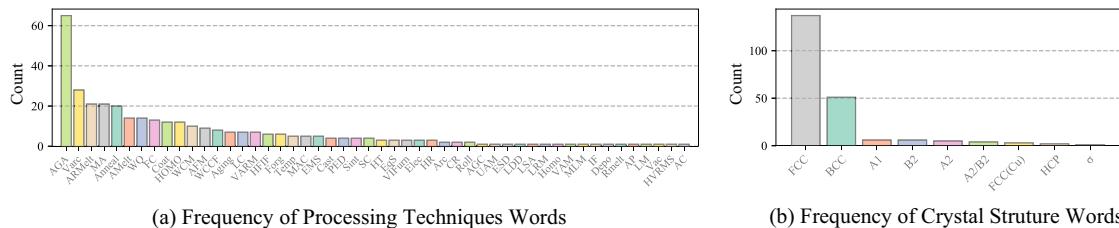


**Fig. 1 | Distribution of elemental content.** This figure shows the distribution of elemental content (in atomic percent, a.t. %) for six elements: Al, Fe, Co, Ni, Cr, Cu, and Mn. The violin plots illustrate the range and distribution of each element across the dataset, with the central box indicating the interquartile range (IQR) and the black line representing the median. The individual data points are shown as dots. The x-axis represents the different elements, while the y-axis corresponds to the atomic percentage of each element.



**Fig. 2 | Correlation analysis between elemental content and corrosion resistance.** Density estimated using Gaussian kernel density estimation. This figure shows the correlation between elemental content (in atomic percent, a.t. %) and corrosion resistance (represented by the logarithm of the corrected current,  $\ln(I_{\text{corr}})$ ). Each panel presents a scatter plot for a specific element (Al, Fe, Co, Ni, Cr, Cu, Mn), with

the x-axis corresponding to the elemental content and the y-axis to the logarithm of the corrosion current. A linear regression line is included to indicate the trend of the correlation. The color density, estimated using Gaussian kernel density estimation, represents the concentration of data points at specific locations within each plot. Correlation coefficients are displayed in each panel.



**Fig. 3 | Frequency distribution of processing parameters and crystal structure information.** This figure shows the frequency distribution of processing technique and crystal structure-related terms. a displays the frequency of various processing technique words, with the x-axis representing the terms and the y-axis showing the count of their occurrences. The colors of the bars represent different categories of

processing techniques. Abbreviations used in the figure are listed in the Supplementary Material, Table S1. **b** shows the frequency of crystal structure-related terms, with the x-axis representing crystal structure types and the y-axis showing the occurrence counts.

elements show weak correlations with  $\ln(I_{\text{corr}})$ . It should be noted that compositional data derived from literature sources generally do not include detailed quantification of trace impurities. Thus, nominally reported zero concentrations may not reflect absolute absence, potentially introducing uncertainty into correlation analyses and model interpretations. The weak correlations, coupled with uncertainties arising from data quality, can complicate model convergence and reduce predictive precision, posing a challenge for the ML process.

Figure 3 presents the distribution of different processing techniques and crystal structures in the HEA-CRD dataset, the distribution of literature on processing techniques in HEA research is uneven, leading to a pronounced long-tail distribution in the data. The terms and phrases in the dataset were directly extracted from the original sources. To avoid improper modifications of the original content and considering the automation requirements of the ML extraction process, no denoising, imputation, or alignment was performed on the extracted data. Due to the lack of standardized descriptions for processing techniques and crystal structures in the literature<sup>52</sup>, the dataset contains noise, coreference issues<sup>53</sup>, and missing information in these descriptions.

Overall, HEA-CRD does not impose restrictions on the processing techniques of HEAs, allowing for a larger dataset and greater diversity in processing methods. Training ML models using HEA-CRD facilitates the construction of intricate relationships among composition, processing, and corrosion resistance, contributing to the broader application of ML in engineering contexts by improving model generalizability and predictive precision. However, this diversity, along with the complexity of element distribution, poor correlation between elements and performance, and the variability in process descriptions, presents challenges for ML modeling<sup>54</sup>.

## Evaluation of the precision of models

We implemented three frameworks using scikit-learn library<sup>55</sup>, with Random Forest (RF) and Multi-Layer Perceptron (MLP) as base models, and developed Mat-NRKG using the PyTorch library<sup>56</sup>. To evaluate the contribution of the structure prediction module in Mat-NRKG, we developed Mat-NRKG<sub>CPP</sub>, in which this module was omitted for comparison. All codes were implemented in Python. To enhance the reliability of experiments with small sample sizes, the data were split into training, validation, and test sets in a 4:1:1 ratio. Six random splits were performed, and the statistical results from these six experiments are reported.

The comparison of experimental results is shown in Table 1. Based on the MSE, MAE, and  $R^2$  metrics, the results from the RF and MLP models across three frameworks indicate that the CPSP framework outperforms the CPP framework, which in turn performs better than the CP framework. This suggests that incorporating processing information and predicted crystal structure data enhances the model's performance in predicting the corrosion resistance of HEAs, indirectly supporting the hypothesis that corrosion resistance is influenced by the combined effects of composition, processing, and crystal structure. Regarding the base models, RF generally performs better than the MLP model. A possible reason for this is that RF, as an ensemble learning method, exhibits stronger generalization capabilities

**Table 1 | Comparative Experimental Results (Testing Set)**

Methods	MSE $\downarrow$	MAE $\downarrow$	$R^2\uparrow$
	Ave. $\pm$ Std.	Ave. $\pm$ Std.	Ave. $\pm$ Std.
RF <sub>CP</sub>	3.06 $\pm$ 0.68	1.41 $\pm$ 0.17	0.35 $\pm$ 0.13
RF <sub>CPP</sub>	2.47 $\pm$ 0.23	1.24 $\pm$ 0.07	0.48 $\pm$ 0.04
RF <sub>CPSP</sub>	2.40 $\pm$ 0.27	1.21 $\pm$ 0.07	0.49 $\pm$ 0.03
MLP <sub>CP</sub>	4.86 $\pm$ 0.49	1.87 $\pm$ 0.14	-0.02 $\pm$ 0.13
MLP <sub>CPP</sub>	3.39 $\pm$ 0.85	1.41 $\pm$ 0.22	0.28 $\pm$ 0.19
MLP <sub>CPSP</sub>	2.91 $\pm$ 0.54	1.32 $\pm$ 0.16	0.39 $\pm$ 0.10
Mat-NRKG <sub>CPP</sub>	2.18 $\pm$ 0.20	1.16 $\pm$ 0.06	0.54 $\pm$ 0.04
<b>Mat-NRKG</b>	<b>1.80 <math>\pm</math> 0.23</b>	<b>1.02 <math>\pm</math> 0.05</b>	<b>0.62 <math>\pm</math> 0.02</b>

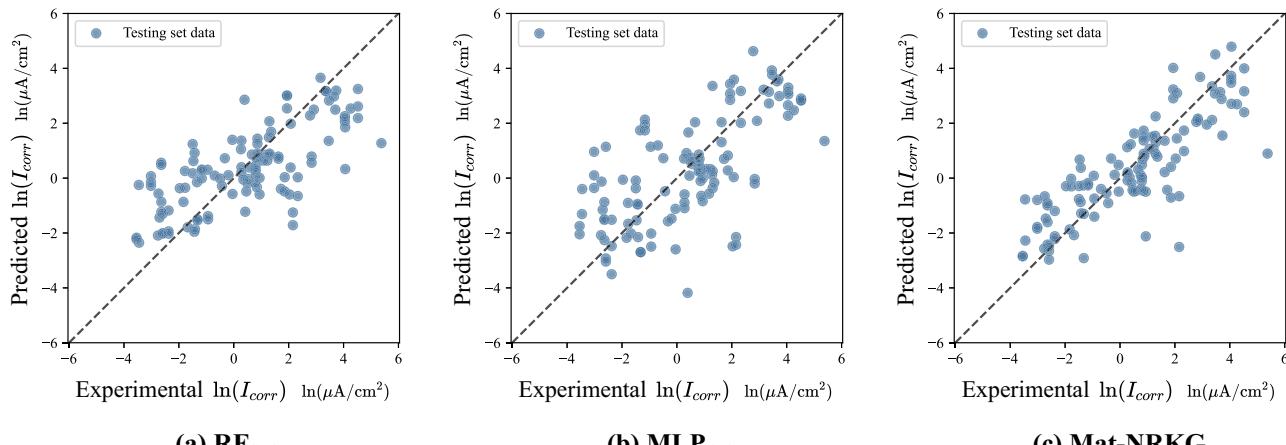
The results are reported as “mean  $\pm$  standard deviation”. The best results are highlighted in bold.

in small-sample scenarios, effectively handling noise and complexity in the data. Compared to RF<sub>CPP</sub>, the inclusion of structural information in RF<sub>CPSP</sub> leads to a 3.1% improvement in  $R^2$ . For the MLP-based models, the improvement is more pronounced: MLP<sub>CPSP</sub> improves  $R^2$  by 35.3% over MLP<sub>CPP</sub>. These results quantitatively shows the benefit of incorporating crystal structure information into the corrosion resistance prediction process.

In contrast, the Mat-NRKG model performs the best among all the compared models, particularly achieving a 25% improvement in the MSE metric over the best-performing comparison model (RF<sub>CPSP</sub>). This improvement further validates the effectiveness of integrating composition, processing information, and predicted crystal structure. Compared to its ablated version Mat-NRKG<sub>CPP</sub>, incorporating structure prediction in Mat-NRKG improves  $R^2$  by 15.1%, showing its effectiveness in enhancing predictive precision. Additionally, the Mat-NRKG<sub>CPP</sub> model still performs slightly worse than the Mat-NRKG but surpasses all comparison models in precision. This indicates that the application of knowledge graphs and GCN-DTB module effectively combines numerical and semantic modalities, establishing correlations between the data and enhancing prediction performance in small-sample scenarios.

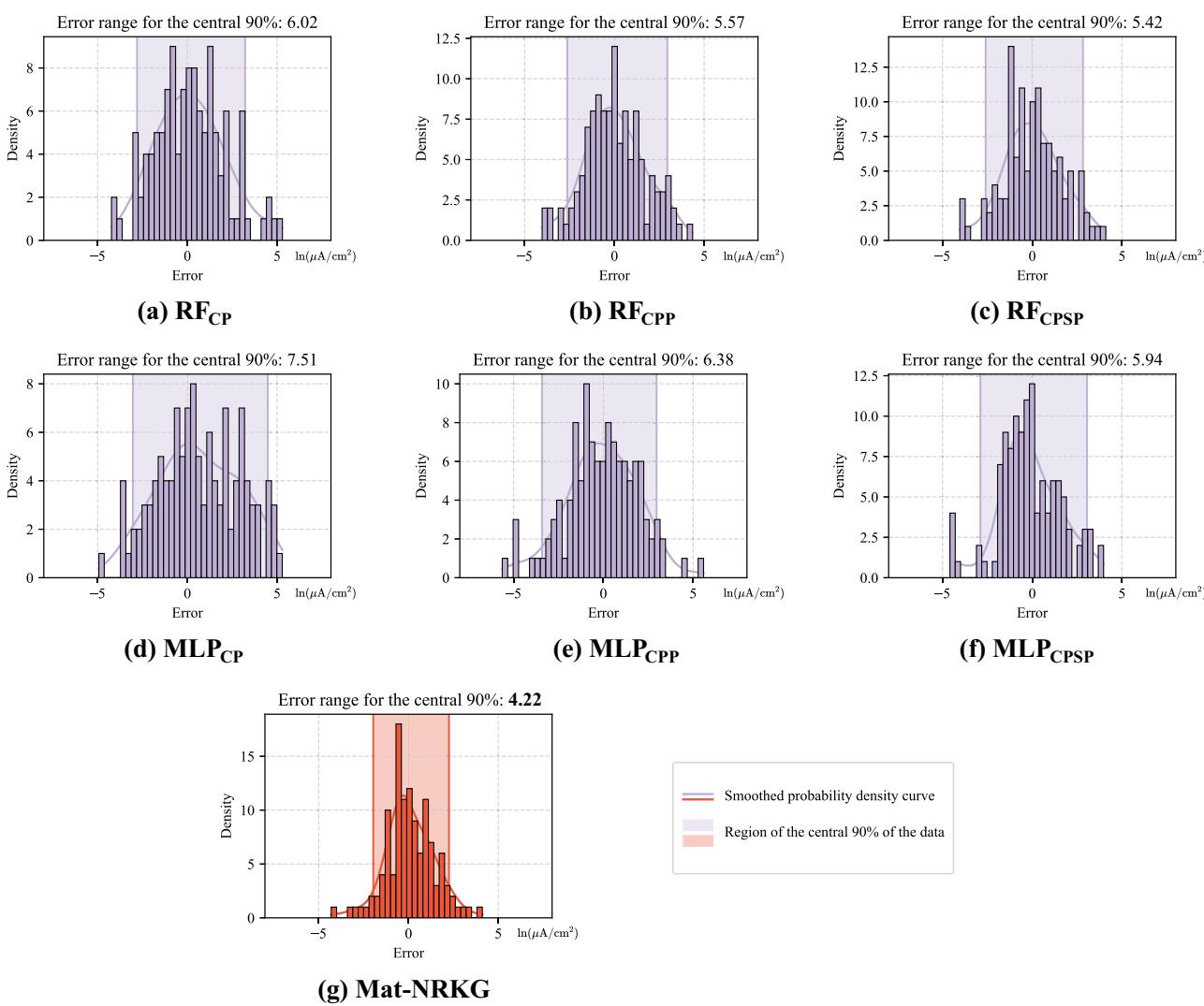
Figure 4 compares the predicted  $\ln(I_{corr})$  from the crystal structure prediction-based CPSP framework models and the Mat-NRKG model with experimental  $\ln(I_{corr})$  for the test set data. The results show that the predicted points from the Mat-NRKG model are more concentrated near the ideal prediction line, indicating superior performance in predicting corrosion current compared to the CPSP framework models.

Figure 5 shows histograms of the absolute error distributions for all test data, with the range covering the central 90% of the errors indicated. As shown, with the progressive inclusion of relevant information, the error distributions for the CP, CPP, and CPSP frameworks become more concentrated, indicating that the prediction errors for most data points decrease as more information is incorporated. Notably, the error distribution for the Mat-NRKG model is the most concentrated, with a shape approaching a



**Fig. 4 | Comparison of predicted and experimental values.** Plots comparing predicted and experimental values of  $\ln(I_{corr})$  on the testing set using three different models: **a** RF<sub>CPSp</sub>, **b** MLP<sub>CPSp</sub>, and **c** Mat-NRKG. Each blue dot represents a data point from the testing set. The dashed diagonal line indicates the ideal case where

predicted values exactly match experimental values. Points closer to the diagonal represent better predictive performance. Mat-NRKG shows a stronger alignment with the ideal line, suggesting improved predictive precision over RF<sub>CPSp</sub> and MLP<sub>CPSp</sub>.



**Fig. 5 | Distribution of Prediction Errors.** This figure shows the distribution of prediction errors for various models: **a** RF<sub>Cp</sub>, **b** RF<sub>Cpp</sub>, **c** RF<sub>Cpsp</sub>, **d** MLP<sub>Cp</sub>, **e** MLP<sub>Cpp</sub>, **f** MLP<sub>Cpsp</sub>, and **g** Mat-NRKG. The histograms illustrate the errors of each model, with the x-axis representing the errors in  $\mu\text{A}/\text{cm}^2$  and the y-axis showing the density.

The shaded regions indicate the central 90% of the data for each distribution. A smoothed probability density curve is overlaid on each histogram for better visualization of the distribution shape. The error range for the central 90% is listed for each model in the top of the respective plots.

normal distribution, further confirming the reliability and stability in predictions.

### Evaluation of the precision of structure prediction

To evaluate the modeling mechanisms of the CPSP framework and the Mat-NRKG model, this study assesses the crystal structure prediction as an intermediate stage in the overall framework

As shown in Table 2, for the crystal structure prediction task, the Mat-NRKG performs better than the CPSP framework in MRR, MR, and Hit@1 metrics. This demonstrates that the graph structure of the knowledge graph effectively captures and integrates unstructured process information, providing a advantage in crystal structure prediction. Additionally, all models achieve a Hit@1 score above 85%, indicating that the CPSP framework also yields promising results in crystal structure prediction task. Although both the CPP and CPSP frameworks take composition and processing information as inputs, the CPSP framework introduces an intermediate crystal structure prediction step, incorporating this structural information into subsequent property predictions. Embedding qualitative domain knowledge into data-driven models is inherently challenging, as such knowledge often involves complex causal relationships that are difficult to quantify. By leveraging the knowledge graph, the CPSP framework integrates explicit instances of how composition and processing influence corrosion resistance through modifications in crystal structure. This structured representation enables the model to learn not only the correlation but also the underlying mechanistic pathways linking material feature variables to performance. As a result, the model demonstrates improved performance on small-sample datasets, better alignment with real-world physical phenomena, and enhanced interpretability.

### Validation of the generalizability of models

To further validate the model's generalization ability, electrochemical tests were conducted on five HEAs synthesized in our laboratory using a casting process. Unlike the literature dataset used for model training, these samples were specifically prepared to provide independent experimental data, enabling an assessment of the model's performance beyond the original dataset. The electrochemical measurements followed the procedures outlined in Sec. 3.8. The material composition, polarization curves, and electrochemical data extracted from the curves are presented in Table 3 and Fig. 6.

To assess the suitability of the synthesized HEAs for generalization testing, we performed a t-SNE analysis on the compositional data. As shown in Fig. 7, the synthesized samples form a distinct cluster that deviates from the main distribution of the HEA-CRD dataset. This indicates that they

represent unseen regions in the composition space, making them appropriate for evaluating the model's generalization capability.

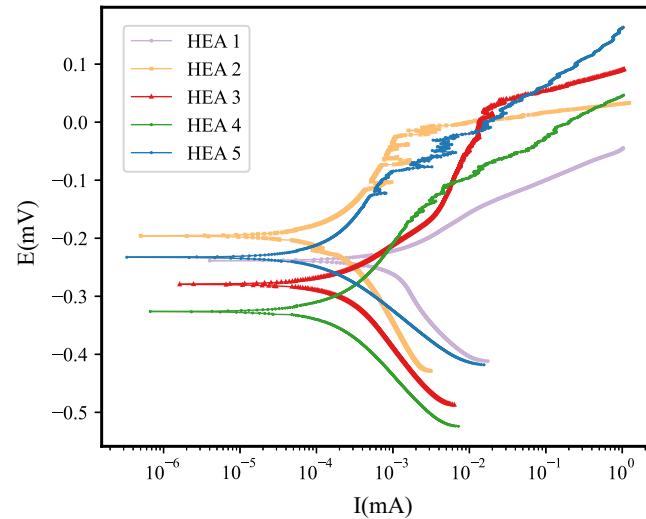
Using the models trained in Sec. 2.2, the synthesized HEAs were tested, and the results are shown in Table 4. The CP framework, CPP framework, and CPSP framework demonstrate progressively improved predictive performance, with Mat-NRKG surpassing the other models, showing generalization ability.

Figure 8 presents a comparison of the CPSP framework models and the Mat-NRKG model for the predicted and experimental values of synthesized HEAs. The predicted values of the HEAs fall within the error range observed for both models on the test set, suggesting that both CPSP and Mat-NRKG models exhibit some level of generalization ability. Notably, the Mat-NRKG model shows good consistency with the testing on HEA-CRD, providing more accurate predictions of corrosion current densities for five different HEAs, and better reflecting the differences in corrosion resistance among the alloys, thus demonstrating superior performance.

Computational resource requirements are provided in the Supplementary Material, see Table S2.

### Interpretability analysis

The CPSP framework and Mat-NRKG enhance interpretability by incorporating crystal structure as a physically meaningful intermediate variable that reflects part of the underlying material mechanisms. As shown in Table 1 and Table 2, the models with higher crystal structure prediction precision tend to yield better performance in corrosion resistance prediction, indicating that the inclusion of structure as an intermediate variable contributes to the model's prediction process.



**Fig. 6 | Polarization curves of synthesized HEAs.** This figure presents the polarization curves for five synthesized HEAs, showing their electrochemical properties. The x-axis shows the corresponding current density ( $I$ ) in millamps, and the y-axis represents the potential ( $E$ ) in millivolts. Each curve corresponds to a different HEA: HEA 1 (purple), HEA 2 (orange), HEA 3 (red), HEA 4 (green), and HEA 5 (blue).

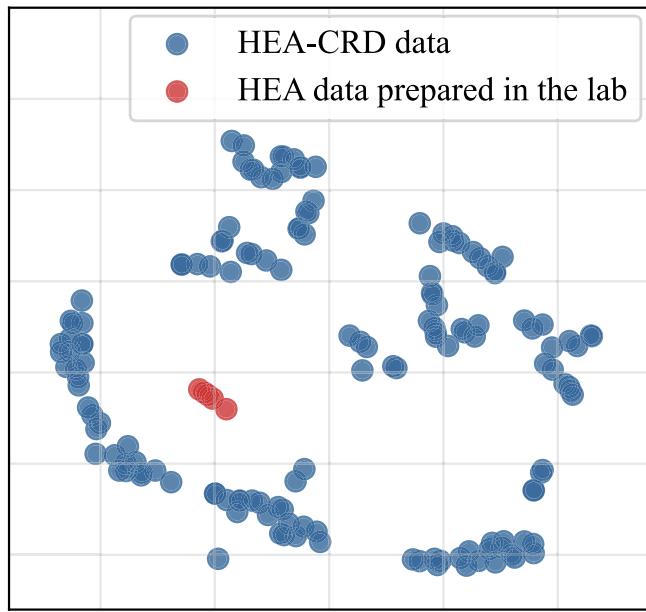
**Table 2 | Crystal Structure Predictions (Testing Set)**

Methods	MRR $\uparrow$ Ave. $\pm$ Std.	MR $\downarrow$ Ave. $\pm$ Std.	Hit@1 $\uparrow$ Ave. $\pm$ Std.
	RF <sub>CPSP</sub>	$0.913 \pm 0.052$	$1.470 \pm 0.319$
MLP <sub>CPSP</sub>	$0.910 \pm 0.052$	$1.331 \pm 0.215$	$0.859 \pm 0.075$
Mat-NRKG	<b><math>0.951 \pm 0.028</math></b>	<b><math>1.170 \pm 0.044</math></b>	<b><math>0.926 \pm 0.013</math></b>

The results are reported as "mean  $\pm$  standard deviation". The best results are highlighted in bold.

**Table 3 | Parameters and Electrochemical Test Results for Synthesized HEAs**

ID	Fe (at%)	Al (at%)	Co (at%)	Cr (at%)	Ni (at%)	Cu (at%)	Mn (at%)	$I_{corr}$ ( $\mu\text{A}/\text{cm}^2$ )	$E_{corr}$ (mV)	$\beta_c$ (mV)	$\beta_a$ (mV)
1	32	0	16	20	6	22	4	0.954	-238.9	229.2	81.6
2	20	0	18	20	22	0	20	0.138	-195.8	149.4	159.4
3	20	0	2	24	26	0	28	0.225	-278.8	165.5	98.2
4	20	0	10	24	26	0	20	0.173	-325.4	145.3	152.3
5	32	0	16	24	20	0	8	0.137	-232.6	106.2	180.1



**Fig. 7 | t-SNE visualization of compositional distribution for HEA Data.** This figure presents a t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization of the compositional distribution for High Entropy Alloy (HEA) data. The blue dots represent data points from HEA-CRD, while the red dot indicates a data point from HEA prepared in the lab. It shows how the data points from different sources are distributed in the compositional space and capture the similarity between data points. The clustering of data points suggests the similarity in compositional characteristics within each group.

**Table 4 | Comparative Experimental Results (Synthesized HEAs)**

Methods	MSE <sub>↓</sub> Ave.±Std.	MAE <sub>↓</sub> Ave.±Std.
RF <sub>CP</sub>	13.27 ± 3.76	3.50 ± 0.53
RF <sub>CPP</sub>	9.22 ± 5.40	2.80 ± 0.90
RF <sub>CPSP</sub>	5.91 ± 1.38	2.31 ± 0.30
MLP <sub>CP</sub>	7.04 ± 1.59	2.48 ± 0.31
MLP <sub>CPP</sub>	6.97 ± 1.57	2.33 ± 0.38
MLP <sub>CPSP</sub>	6.06 ± 0.96	2.32 ± 0.20
<b>Mat-NRKG</b>	<b>2.64 ± 1.66</b>	<b>1.37 ± 0.57</b>

The results are reported as “mean ± standard deviation”. The best results are highlighted in bold.

To further evaluate interpretability in the Mat-NRKG model, the learned embeddings of processing technique nodes were analyzed, and their pairwise similarities were visualized. As shown in Fig. 9, these embeddings were initialized randomly and trained without any explicit semantic labeling. After training, several meaningful similarities emerged. For example, processes such as laser re-melting, laser melting, laser cladding, and laser direct deposition showed strong similarity, corresponding to shared laser-based mechanisms. Other groups included coating-related techniques, casting-related operations such as argon gas atmosphere, arc re-melting, vacuum arc, arc furnace melting, casting, and argon gas casting, as well as cooling methods like furnace cooling, slow cooling, and water-cooled copper crucible furnace. These associations are consistent with established domain knowledge and suggest that the model has internally captured relevant relationships among processing techniques. Importantly, such learned associations help mitigate the “coreference issue”, where semantically identical or similar processing terms are written differently in the literature. By embedding these variations into a continuous representation space, the model reduces redundancy and noise caused by inconsistent

terminology. Such representations may provide additional context for predicting material properties and support the model’s ability to generalize from diverse input data.

### Significance and limitations of the framework

CPSp framework illustrates approaches to embedding domain knowledge—such as the intricate relationships between composition, processing, and crystal structure—into ML models. Capturing these dependencies within an ML framework is inherently challenging due to the qualitative nature of materials science knowledge and the complexity of composition-structure-processing-performance relationships. By leveraging diverse data and integrating structured prior knowledge, these frameworks establish a more unified and consistent predictive model, enhancing the interpretability and reliability of predictions.

Building upon this foundation, this study proposes the Mat-NRKG method, which advances this integration by explicitly utilizing a knowledge graph to encode composition and processing information. By aligning these features with crystal structure prediction, Mat-NRKG enables a more structured learning paradigm, improving the model’s ability to forecast HEA corrosion resistance. Compared to existing models, the Mat-NRKG model demonstrates superior predictive precision in evaluating HEA corrosion resistance, providing value for the rapid screening of materials in engineering applications.

It should be noted that the models are trained and evaluated using the logarithm of corrosion current density,  $\ln(I_{\text{corr}})$ , which is a common practice in corrosion science due to the wide dynamic range and its alignment with electrochemical kinetics. However, even small deviations on the  $\ln$ -scale may correspond to large absolute errors in linear scale, potentially impacting alloy screening in practical applications. Therefore, future work may consider incorporating dual-scale error control or calibration schemes for practical use.

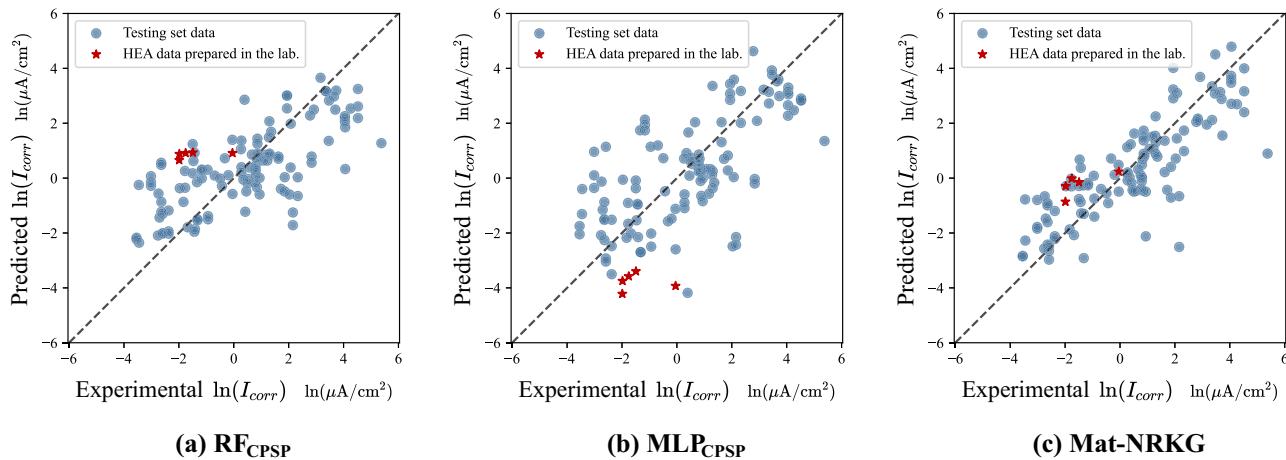
Currently, the proposed models are limited to HEA corrosion resistance data measured under conditions of 25 °C (or room temperature) and 3.5 wt% NaCl solution. Future work will focus on developing a dataset that includes corrosion resistance data under various experimental conditions. Additionally, the use of hyper-relational knowledge graphs will be explored to model data under different testing conditions, facilitating the development of an end-to-end predictive model that integrates composition, processing, and experimental conditions to further improve the model’s applicability and prediction precision.

## Methods

### Overview of HEA corrosion resistance prediction frameworks

The Composition and Processing-Driven Two-Stage Corrosion Prediction Framework (CPSp Framework, Fig. 10c, details in Sec. 3.5) is proposed to predict the logarithm of corrosion current density  $\ln(I_{\text{corr}})$  of HEAs. This framework first predicts the crystal structure based on composition and processing information, and subsequently utilizes the predicted structure along with composition and processing conditions to estimate corrosion current. Building upon the conceptual foundation of CPSp, the NRKG-S method (Fig. 11, details in Sec. 3.6) is further developed, integrating knowledge graph technology<sup>45</sup> to enhance predictive precision and improve model interoperability.

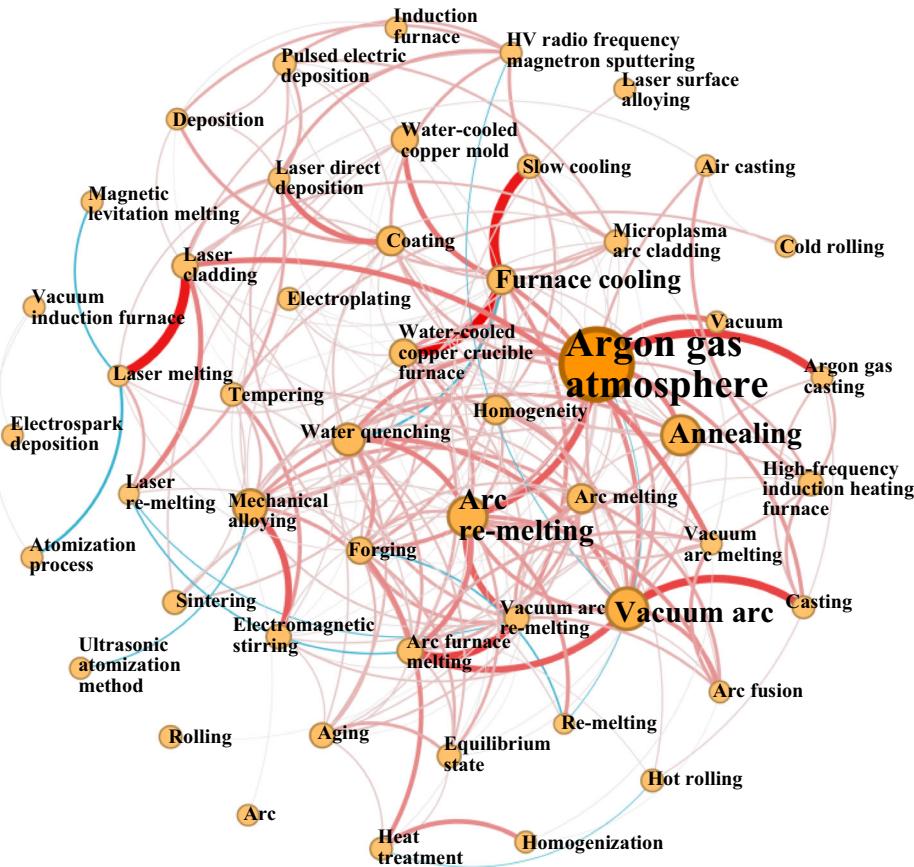
To establish comparative baselines, two conventional frameworks are designed: the Composition-Only Prediction Framework (CP Framework, details in Sec. 3.3), which estimates corrosion current based solely on composition, and the Composition and Processing-Based Prediction Framework (CPP Framework, details in Sec. 3.4), which incorporates both composition and processing conditions as input features. The CP, CPP, and CPSp frameworks allow flexibility in the selection of ML models as base model, theoretically accommodating most regression and classification methods. For implementation, Random Forest (RF)<sup>57</sup> and Multilayer Perceptron (MLP)<sup>58</sup> are chosen as the base models (details in Sec. 3.2), given their widespread application in materials property prediction. The subsequent sections provide a comprehensive description of each framework and method.



**Fig. 8 | Comparison of predicted and experimental values for synthesized HEAs.** Plots comparing predicted and experimental values of on the testing set using three different models: a RF<sub>CPSp</sub>, b MLP<sub>CPSp</sub>, and (c) Mat-NRKG. Each blue dot represents a data point from the testing set, while the red stars indicate data points for HEAs prepared in the lab. The dashed diagonal line represents the ideal case where the

predicted values match exactly with the experimental values. Points that are closer to the diagonal line represent better predictive performance. Among the three models, Mat-NRKG exhibits a stronger alignment with the ideal line, indicating better predictive precision compared to RF<sub>CPSp</sub> and MLP<sub>CPSp</sub>.

**Fig. 9 | Embedding similarity among processing technique nodes.** This figure visualizes the embedding similarity between processing technique nodes in a canonical knowledge graph. The nodes represent various processing techniques, with node size proportional to the frequency of occurrence of the techniques. The edges between the nodes indicate cosine similarity in their embedding space, with edge thickness reflecting the magnitude of the similarity. Red edges represent positive similarity, while blue edges correspond to negative similarity.



### Base model

The Random Forest (RF) model<sup>57</sup> is an ensemble learning method that constructs multiple decision trees on randomly sampled data subsets. The final prediction is obtained by averaging individual tree outputs:

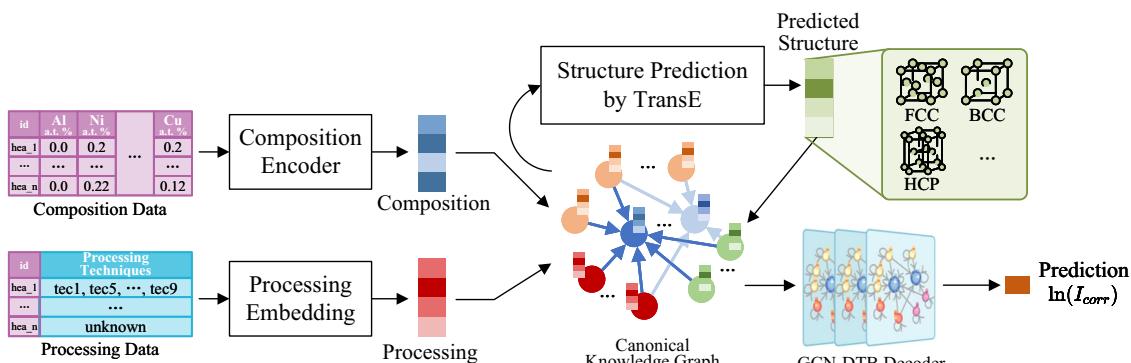
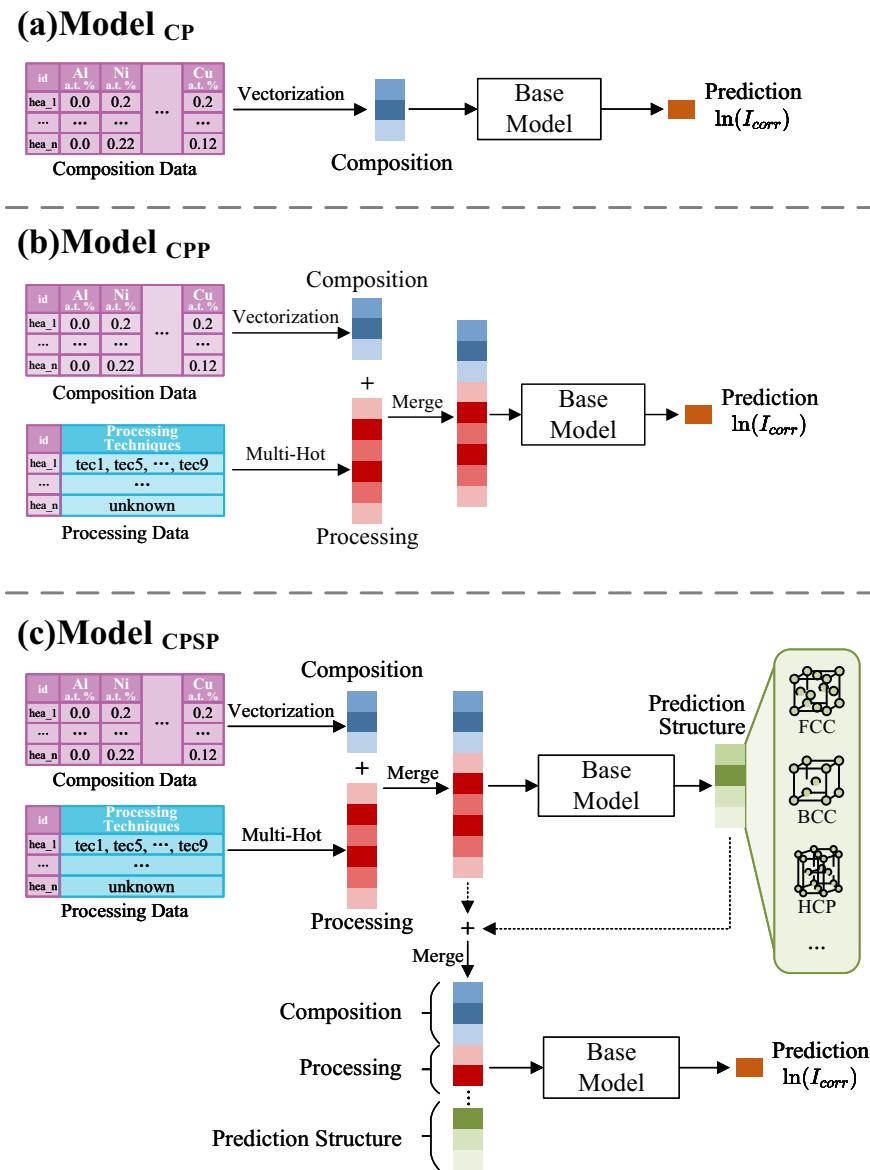
$$\hat{y} = \frac{1}{n} \sum_{i=1}^n \text{Tree}_i(x) \quad (1)$$

where  $n$  is the number of trees and  $x$  is the input feature vector. RF is effective for high-dimensional, noisy data due to its robustness and interpretability.

The Multilayer Perceptron (MLP)<sup>58</sup> is a feedforward neural network that learns complex feature interactions through multiple nonlinear transformations:

$$\mathbf{h}^{(l)} = f(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}) \quad (2)$$

**Fig. 10 | HEA corrosion resistance prediction frameworks.** This figure presents three different frameworks for predicting the corrosion resistance of High Entropy Alloys. **a** Composition-Only Prediction Framework (CP Framework): The framework incorporates composition data, which is vectorized and fed into a base model to predict the logarithmic corrosion rate  $\ln(I_{corr})$ . **b** Composition and Processing-Based Prediction Framework (CPP Framework): In addition to the composition data, this model integrates processing data. Processing techniques are represented as multi-hot encoded vectors, and the composition and processing data are merged before being passed to the base model for prediction. **c** Composition and Processing-Driven Two-Stage Corrosion Prediction Framework with Structural Prediction (CPSP Framework): This model further extends the framework by incorporating a prediction structure that accounts for crystal structure. Composition and processing data are merged, and the model predicts the corrosion resistance based on this extended input. Each model progressively adds more variables to improve the prediction of HEA corrosion resistance, with the final output being the predicted logarithmic corrosion rate  $\ln(I_{corr})$ .



**Fig. 11 | HEA corrosion resistance prediction method mat-NRKG.** This figure shows the Mat-NRKG method for predicting HEA corrosion resistance. It integrates composition and processing data through separate encoders, uses a knowledge

graph to predict crystal structures with the TransE algorithm, and then predicts the corrosion current ( $\ln(I_{corr})$ ) using a GCN-DTB decoder.

where,  $\mathbf{h}^{(l)}$  is the output of layer  $l$ , and  $f(\cdot)$  is the activation function. MLP excels in capturing nonlinear patterns, making it suitable for intricate prediction tasks.

### Composition-only prediction framework

As shown in Fig. 10a, the Composition-Only Prediction Framework (CP Framework) directly predicts the corrosion current  $\ln(I_{\text{corr}})$  using only the compositional information of HEAs. The atomic percentages of Al, Fe, Co, Ni, Cr, Cu, and Mn in the HEA composition are represented as a 7-dimensional vector (compositional features), which is normalized before being input into the base model for training and prediction. In this study, RF and MLP are employed as the base models. Serving as a baseline, the CP Framework provides a fundamental assessment of the predictive capability of compositional features alone, facilitating comparison with more advanced frameworks.

### Composition and processing-based prediction framework

As shown in Fig. 10b, the Composition and Processing-Based Prediction Framework (CPP Framework) integrates compositional and processing information of HEAs to enhance the prediction of corrosion current  $\ln(I_{\text{corr}})$ . Processing information, described by phrases, cannot be directly input into the model due to the variability in the number of processing associated with each HEA. To address this, all possible processing conditions are listed as a feature set. For each HEA, a fixed-length vector is maintained, where each position corresponds to a specific processing condition. If a particular processing method is applied, the corresponding position in the vector is marked as “1”; otherwise, it is marked as “0”. This approach constitutes a multi-hot encoding scheme<sup>59</sup>, enabling the transformation of categorical processing information into a structured numerical format suitable for ML models. The advantage of multi-hot encoding lies in its ability to represent multiple processing conditions simultaneously while maintaining a consistent input size across all samples. After concatenating compositional features (consistent with Sec. 3.3) and processing encodings, the combined input is fed into the base model for training and prediction. By incorporating processing information, the model captures the relationships between composition, processing, and corrosion current. In this study, the base model is implemented using RF and MLP. As an extension beyond the CP Framework, the CPP Framework serves as another baseline, allowing evaluation of the impact of both compositional and processing information on corrosion prediction.

### Composition and processing-driven two-stage corrosion prediction framework with structural prediction

Given that processing can influence the crystal structure, which in turn affects corrosion resistance, yet directly including structure as an input parameter limits the model's engineering applicability due to the need for experimental or computational determination, Fig. 10c shows the Composition and Processing-Driven Two-Stage Corrosion Prediction Framework with Structural Prediction (CPSP Framework). In the first stage, the framework takes compositional and processing information as inputs (consistent with Sec. 3.3) and uses a multiclass classifier to predict the crystal structure type of the multi-hot encoded material. Crystal structure labels extracted from the literature are used to supervise the training of this stage, and also serve as ground truth for evaluating prediction precision during testing. In the second stage, compositional and processing features, along with the predicted crystal structure from the first stage, are concatenated and input into the base model to further predict corrosion current. This method, by introducing the intermediate variable predicted crystal structure, may be able to capture the indirect effects of composition and processing on corrosion resistance performance. The first stage is a multiclass classification task, and the base model selected is a classifier such as RF or MLP. To ensure comparability, the same base model (RF or MLP) is used in both stages of the framework.

### Mat-NRKG

While the CPSP framework incorporates a physically meaningful intermediate step by introducing crystal structure prediction, which reflects how

composition and processing affect material behavior, its two-stage design results in a gap between the first stage and corrosion resistance performance. To bridge this gap and further build upon the CPSP framework, inspired by Song et al.<sup>45</sup> the Mat-NRKG method is proposed. Additionally, the compositional complexity of HEAs, along with the impact of composition and processing techniques on crystal structure and corrosion resistance, poses challenges for accurate property prediction. The numerical reasoning method for material knowledge graphs (NR-KG)<sup>45</sup> was the first to construct a cross-modal knowledge graph, integrating heterogeneous numerical and semantic representations of composition, processing, and crystal structure into a unified modeling framework. However, NR-KG relies on crystal structure as an explicit input, limiting its applicability in scenarios where structural data is impractical to obtain through experiments or simulations. To address this limitation, the CPSP framework's concept of predicting crystal structure is leveraged to enhance NR-KG, allowing structure prediction from composition and processing information while maintaining a unified knowledge-driven reasoning process. This modification has the potential to improve model generalizability and reduce dependency on explicit structural data, thereby potentially enhancing its applicability in engineering applications for corrosion resistance prediction.

As shown in Fig. 11, the canonical knowledge graph encodes numerical (composition) and semantic (processing, crystal structure) data in a graph structure. Specifically, the composition, processing, and candidate crystal structures of HEAs are modeled as nodes in the knowledge graph. The relationships between nodes are constructed based on the compositional ratios, processing techniques, and their potential impact on crystal structure. Compared to multi-hot encoding method, the topological structure of the knowledge graph allows for a more flexible capture of multidimensional semantic information<sup>46</sup>, especially in cases where processing techniques are complex and variable. Each node in the canonical knowledge graph records relevant feature information: composition data is extracted by the Composition Encoder, which uses an MLP-based embedding network to map the 7-dimensional atomic percentage vector to a low-dimensional feature vector. Specifically, this process effectively performs an embedding transformation, projecting the original composition data into a continuous vector space while aligning it with the dimensionality of node embeddings. This transformation enhances feature expressiveness and facilitates integration with the overall knowledge graph representation. Processing data is handled by the Processing Embedding module, which uses an embedding lookup method<sup>60</sup> to learn optimizable embeddings for each processing parameter. Specifically, since processing nodes are inherently discrete, this module follows a strategy inspired by natural language embedding techniques: it first initializes feature embeddings randomly and then refines them through iterative optimization during the training process. This approach ensures that the learned embeddings capture meaningful relationships among different processing techniques, improving the knowledge graph's ability to encode process-related semantics.

The crystal structure prediction task for HEAs is modeled as a link prediction task within the canonical knowledge graph, where the goal is to determine the probability of a connection between an HEA node and a candidate crystal structure node. Specifically, this involves assessing the likelihood that a given HEA composition adopts a particular crystal structure, based on the relationships encoded in the knowledge graph. To achieve this, the TransE algorithm<sup>47</sup> is employed to model these connections and predict potential structural configurations. For a triplet in the knowledge graph,  $(h, r, t)$ , where  $h$  is the head node,  $r$  is the relationship, and  $t$  is the tail node, the TransE algorithm optimizes the following objective function:

$$L = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2 \quad (3)$$

where  $\mathbf{h}$ ,  $\mathbf{r}$ , and  $\mathbf{t}$  are the embedding vectors for nodes  $h$ , relationship  $r$ , and node  $t$ , respectively, and  $\|\cdot\|_2$  denotes the  $L_2$  norm. By minimizing the distance  $L$ , the TransE algorithm effectively integrates both node features (such as composition information) and the knowledge graph's topological structure (which encodes the semantic similarity between processing

techniques). This enables the model to capture the potential influence of composition and processing parameters on the resulting crystal structure, thereby enhancing the model's ability to reproduce and learn complex structure-related patterns, which contributes to improved predictive performance in downstream tasks such as corrosion resistance prediction.

After the structure prediction, the model further utilizes a GCN-DTB module, which combines a Graph Convolutional Network (GCN)<sup>48</sup> decoder with a Deep Taylor Block (DTB) module<sup>49</sup>, to aggregate multi-layered information from the HEA node and its neighboring nodes (processing and structure nodes) in the canonical knowledge graph. This captures the complex effects of composition, processing, and structure on corrosion performance to predict the corrosion current. The GCN-based prediction expression is given by:

$$\hat{y} = \sigma \left( W^{(l)} \cdot \sum_{u \in \mathcal{N}(v)} \frac{1}{c_{uv}} \cdot h_u^{(l-1)} + b^{(l)} \right) \quad (4)$$

where  $h_u^{(l-1)}$  represents the feature of the neighboring node  $u$  at layer  $l-1$ ,  $c_{uv}$  is the normalization factor for adjacent nodes,  $W^{(l)}$  is the weight matrix, and  $\sigma$  is the activation function. The DTB module helps reduce the impact of data noise on the modeling process.

Mat-NRKG adopts the core concept of the CPSP framework, where compositional and processing information is first used to predict crystal structure, followed by the integration of all information for corrosion resistance prediction. Compared to CPSP, Mat-NRKG transforms the two-stage prediction process into an end-to-end prediction within the knowledge graph framework. By modeling semantic relationships in the knowledge graph, the framework improves the fusion of composition and processing data and leverages the GCN's neighborhood aggregation mechanism to capture complex multi-factor interactions.

## Evaluation metrics

This study uses the polarization method to measure  $\ln(I_{corr})$  as an indicator of the corrosion resistance of HEAs, which is considered a standard approach in corrosion science for evaluating electrochemical behavior<sup>16,61,62</sup>. Accordingly, we use regression-based metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination ( $R^2$ ) to evaluate model performance, as they are widely adopted in the literature for assessing predictive precision in materials informatics<sup>63,64</sup>. The metrics are defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (7)$$

where  $y_i$  and  $\hat{y}_i$  are the true and predicted values for sample  $i$ ,  $\bar{y}$  is the mean of the true values, and  $N$  is the total number of samples.

Additionally, the CPSP framework and Mat-NRKG method proposed here can predict the crystal structure. To evaluate the effectiveness of this task, we use the Mean Reciprocal Rank (MRR), Mean Rank (MR), and Hit@1, defined as:

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank_i} \quad (8)$$

$$MR = \frac{1}{N} \sum_{i=1}^N rank_i \quad (9)$$

$$Hit@1 = \frac{\text{Correct}@1}{N} \quad (10)$$

where  $rank_i$  denotes the rank of the correct answer for sample  $i$ , and  $\text{Correct}@1$  is the number of correct predictions at rank 1.

## Electrochemical measurements

To validate the model performance, electrochemical measurements were conducted on five laboratory-synthesized HEAs prepared by a casting process. Electrochemical tests were performed using a Gamry Reference 600+ electrochemical workstation. The HEA samples were fabricated via a casting process and subsequently cured at room temperature for 24 h. Prior to testing, the samples were sequentially polished with 400#, 800#, 1600#, and 3000# sandpaper to achieve a uniform surface finish. The samples were then cleaned ultrasonically in anhydrous ethanol for 5 minutes and dried with cold air.

Electrochemical measurements were conducted in a 3.5% NaCl solution at room temperature using a conventional three-electrode setup. The working electrode (WE) was the synthesized HEA, embedded in epoxy resin with a single exposed surface of 1 cm<sup>2</sup>. A saturated calomel electrode (SCE) was used as the reference electrode (RE), and a platinum plate (2 cm × 2 cm) was used as the counter electrode (CE). The SCE was commercially pre-calibrated, and its stability was confirmed by repeated open-circuit potential (OCP) measurements and comparison with a standard redox couple.

Prior to polarization testing, the samples were immersed in the NaCl solution for 4 h to stabilize the surface condition and reach electrochemical equilibrium. The OCP was then monitored, and once it remained within ±2 mV for at least 10 min, the polarization scan was initiated. The potentiodynamic polarization (PDP) test started at -0.2 V relative to OCP and proceeded to +2 V versus the reference electrode, with a scan rate of 0.5 mV/s and a termination current of 1 mA/cm<sup>2</sup>.

The corrosion current density  $I_{corr}$  and corrosion potential  $E_{corr}$  were extracted from the polarization curves using EC-Lab software (Version 9.32). To minimize the influence of passivation behavior observed in the anodic branch, only the cathodic Tafel region was selected for extrapolation. A linear fit was performed on the cathodic branch within the region that exhibited a clear Tafel slope. The  $I_{corr}$  value was obtained by extending the fitted cathodic line to intersect the potential corresponding to the minimum current on the polarization curve.  $E_{corr}$  was determined as the potential at which the anodic and cathodic branches intersected. This cathodic extrapolation method was adopted to improve the precision and stability of the electrochemical parameter extraction in HEAs exhibiting partial anodic passivation.

## Data availability

The data is publicly available at <https://github.com/MatrixBrain/NR-KG/tree/main/dataset>.

## Code availability

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

Received: 12 March 2025; Accepted: 17 June 2025;

Published online: 04 July 2025

## References

1. Tung, C.-C. et al. On the elemental effect of AlCoCrCuFeNi high-entropy alloy system. *Mater. Lett.* **61**, 1–5 (2007).
2. Birbilis, N., Choudhary, S., Scully, J. & Taheri, M. A perspective on corrosion of multi-principal element alloys. *npj Mater. Degrad.* **5**, 14 (2021).
3. Kim, J. H., Lim, K. R., Won, J. W., Na, Y. S. & Kim, H.-S. Mechanical properties and deformation twinning behavior of as-cast CoCrFeMnNi high-entropy alloy at low and high temperatures. *Mater. Sci. Eng.: A* **712**, 108–113 (2018).

4. Bae, J. W. et al. Trade-off between tensile property and formability by partial recrystallization of CrMnFeCoNi high-entropy alloy. *Mater. Sci. Eng.: A* **703**, 324–330 (2017).
5. Kim, Y.-K., Joo, Y.-A., Kim, H. S. & Lee, K.-A. High temperature oxidation behavior of cr-mn-fe-co-ni high entropy alloy. *Intermetallics* **98**, 45–53 (2018).
6. Wang, Y., Zhang, M., Jin, J., Gong, P. & Wang, X. Oxidation behavior of CoCrFeMnNi high entropy alloy after plastic deformation. *Corros. Sci.* **163**, 108285 (2020).
7. Kai, W. et al. The corrosion of an equimolar FeCoNiCrMn high-entropy alloy in various CO<sub>2</sub>/CO mixed gases at 700 and 950°. *Corros. Sci.* **153**, 150–161 (2019).
8. Wang, Y. et al. Influence of plastic deformation on the corrosion behavior of CrCoFeMnNi high entropy alloy. *J. Alloy. Compd.* **891**, 161822 (2022).
9. Wang, Z., Li, D., Yao, Y.-Y., Kuo, Y.-L. & Hsueh, C.-H. Wettability, electron work function and corrosion behavior of CoCrFeMnNi high entropy alloy films. *Surf. Coat. Technol.* **400**, 126222 (2020).
10. Torbati-Sarraf, H., Shabani, M., Jablonski, P. D., Pataky, G. J. & Poursaei, A. The influence of incorporation of mn on the pitting corrosion performance of CrFeCoNi high entropy alloy at different temperatures. *Mater. Des.* **184**, 108170 (2019).
11. Jin, Z. et al. Effect of annealing temperature on the microstructure evolution and corrosion behavior of carbon-interstitial FeMnCoCrNi high-entropy alloys. *Corros. Sci.* **228**, 111813 (2024).
12. Sasidhar, K. N. et al. Deep learning framework for uncovering compositional and environmental contributions to pitting resistance in passivating alloys. *Npj Mater. Degrad.* **6**, 71 (2022).
13. Song, Y. et al. Interpretability study on prediction models for alloy pitting based on ensemble learning. *Corros. Sci.* **228**, 111790 (2024).
14. Sathiyamoorthi, P. & Kim, H. S. High-entropy alloys with heterogeneous microstructure: Processing and mechanical properties. *Prog. Mater. Sci.* **123**, 100709 (2022).
15. Sonar, T., Ivanov, M., Trofimov, E., Tingaev, A. & Suleymanova, I. An overview of microstructure, mechanical properties and processing of high entropy alloys and its future perspectives in aeroengine applications. *Mater. Sci. Energy Technol.* **7**, 35–60 (2024).
16. Fu, Y., Li, J., Luo, H., Du, C. & Li, X. Recent advances on environmental corrosion behavior and mechanism of high-entropy alloys. *J. Mater. Sci. Technol.* **80**, 217–233 (2021).
17. Xu, W., Diesen, E., He, T., Reuter, K. & Margraf, J. T. Discovering high entropy alloy electrocatalysts in vast composition spaces with multiobjective optimization. *J. Am. Chem. Soc.* **146**, 7698–7707 (2024).
18. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
19. Rao, Z. et al. Machine learning–enabled high-entropy alloy discovery. *Science* **378**, 78–85 (2022).
20. Liu, M., Lei, C., Wang, Y., Zhang, B. & Qu, X. High-throughput preparation for alloy composition design in additive manufacturing: a comprehensive review. *Mater. Genome Eng. Adv.* **2**, e55 (2024).
21. Ji, Y. et al. Artificial intelligence combined with high-throughput calculations to improve the corrosion resistance of AlMgZn alloy. *Corros. Sci.* **233**, 112062 (2024).
22. Vitos, L. *Computational quantum mechanics for materials engineers: the EMTO method and applications* (Springer Science & Business Media, 2007).
23. Lee, C. et al. Temperature dependence of elastic and plastic deformation behavior of a refractory high-entropy alloy. *Sci. Adv.* **6**, eaaz4748 (2020).
24. Schleider, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: Recent approaches to materials science—a review. *J. Phys.: Mater.* **2**, 032001 (2019).
25. Yip, S. & Short, M. P. Multiscale materials modelling at the mesoscale. *Nat. Mater.* **12**, 774–777 (2013).
26. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
27. Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018).
28. Maman, O., Wenzlick, M., Sathanur, A., Hawk, J. & Devanathan, R. Machine learning augmented predictive and generative model for rupture life in ferritic and austenitic steels. *Npj Mater. Degrad.* **5**, 20 (2021).
29. Ren, J.-C. et al. Predicting single-phase solid solutions in as-sputtered high entropy alloys: high-throughput screening with machine-learning model. *J. Mater. Sci. Technol.* **138**, 70–79 (2023).
30. Wei, Q. et al. Discovering a formula for the high temperature oxidation behavior of FeCrAlCoNi based high entropy alloys by domain knowledge-guided machine learning. *J. Mater. Sci. Technol.* **149**, 237–246 (2023).
31. Jin, X. et al. Data mining accelerated the design strategy of high-entropy alloys with the largest hardness based on genetic algorithm optimization. *Mater. Genome Eng. Adv.* **2**, e49 (2024).
32. Zhao, X., Huang, H., Su, Y., Qiao, L. & Yan, Y. Exploring high corrosion-resistant refractory high-entropy alloy via a combined experimental and simulation study. *npj Mater. Degrad.* **8**, 77 (2024).
33. Qiao, L., Ramanujan, R. & Zhu, J. Machine learning accelerated design of a family of AlxCrFeNi medium entropy alloys with superior high temperature mechanical and oxidation properties. *Corros. Sci.* **211**, 110805 (2023).
34. Dong, Z. et al. Machine learning-assisted discovery of cr, al-containing high-entropy alloys for high oxidation resistance. *Corros. Sci.* **220**, 111222 (2023).
35. Song, G. et al. Bridging the semantic-numerical gap: a numerical reasoning method of cross-modal knowledge graph for material property prediction. *arXiv preprint arXiv:2312.09744* (2023).
36. Beckh, K. et al. Explainable machine learning with prior knowledge: an overview. *arXiv preprint arXiv:2105.10172* (2021).
37. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Acc. Mater. Res.* **3**, 597–607 (2022).
38. Luo, H. et al. A strong and ductile medium-entropy alloy resists hydrogen embrittlement and corrosion. *Nat. Commun.* **11**, 3081 (2020).
39. Miracle, D. B. et al. Exploration and development of high entropy alloys for structural applications. *Entropy* **16**, 494–525 (2014).
40. Takeuchi, A. et al. Entropies in alloy design for high-entropy and bulk glassy alloys. *Entropy* **15**, 3810–3821 (2013).
41. Yang, X. & Zhang, Y. Prediction of high-entropy stabilized solid-solution in multi-component alloys. *Mater. Chem. Phys.* **132**, 233–238 (2012).
42. Ye, Y., Wang, Q., Lu, J., Liu, C. & Yang, Y. Design of high entropy alloys: A single-parameter thermodynamic rule. *Scr. Mater.* **104**, 53–55 (2015).
43. Gao, M. C. et al. Computational modeling of high-entropy alloys: structures, thermodynamics and elasticity. *J. Mater. Res.* **32**, 3627–3641 (2017).
44. Wu, M., Wang, S., Huang, H., Shu, D. & Sun, B. CALPHAD aided eutectic high-entropy alloy design. *Mater. Lett.* **262**, 127175 (2020).
45. Song, G., Fu, D. & Zhang, D. From knowledge graph development to serving industrial knowledge automation: A review. in 2022 41st Chinese control conference (CCC) 4219–4226 (IEEE, 2022).
46. Hogan, A. et al. Knowledge graphs. *ACM Comput. Surv.* **54**, 1–37 (2021).
47. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Adv. Neural Inform. Process. Syst.* **26**, 1–9 (2013).
48. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)* (2017).

49. Song, G., Fu, D., Qiu, Z., Meng, J. & Zhang, D. Taylor-sensus network: Embracing noise to enlighten uncertainty for scientific data. arXiv preprint arXiv:2409.07942 (2024).
50. Wissler, C. The spearman correlation formula. *Science* **22**, 309–311 (1905).
51. Iversen, G. R. & Gergen, M. (eds.) *Statistics, the Conceptual Approach*. (Springer, New York, 1997).
52. Hill, J. et al. Materials science with large-scale data and informatics: Unlocking new opportunities. *Mrs Bull.* **41**, 399–409 (2016).
53. Peng, H., Khashabi, D. & Roth, D. Solving hard coreference problems. *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.* 809–819 (2015).
54. Kumar, P., Bhatnagar, R., Gaur, K. & Bhatnagar, A. Classification of imbalanced data: review of methods and applications. In IOP conference series: Materials science and engineering vol. 1099 012077 (IOP Publishing, 2021).
55. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Paszke, A. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.* **32**, 8024–8035 (2019).
57. Belgiu, M. & Drăguț, L. Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016).
58. Heidari, A. A., Faris, H., Aljarah, I. & Mirjalili, S. An efficient hybrid multilayer perceptron neural network with grasshopper optimization. *Soft Comput.* **23**, 7941–7958 (2019).
59. Li, B. et al. EMU: Effective multi-hot encoding net for lightweight scene text recognition with a large character set. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 5374–5385 (2022).
60. Church, K. W. Word2Vec. *Nat. Lang. Eng.* **23**, 155–162 (2017).
61. Qiu, Y., Thomas, S., Gibson, M. A., Fraser, H. L. & Birbilis, N. Corrosion of high entropy alloys. *npj Mater. Degrad.* **1**, 15 (2017).
62. Qiu, Y., Gibson, M., Fraser, H. & Birbilis, N. Corrosion characteristics of high entropy alloys. *Mater. Sci. Technol.* **31**, 1235–1243 (2015).
63. Xu, P., Ji, X., Li, M. & Lu, W. Small data machine learning in materials science. *npj Comput. Mater.* **9**, 42 (2023).
64. Hart, G. L., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).

## Acknowledgements

This work was supported by the National Key R&D Program of China (2024YFF0728900) and the National Environmental Corrosion Platform of China (KXJS2023002).

## Author contributions

G.S. conceptualized the study, developed the methodology, conducted data curation, and drafted the original manuscript. D.F. oversaw project administration, provided supervision, and contributed to conceptualization. Y.L. was responsible for resources and data curation. L.M. and D.Z. supervised the study and contributed to writing, review, and editing. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41529-025-00632-4>.

**Correspondence** and requests for materials should be addressed to Guangxuan Song, Dongmei Fu, Yongjie Lin, Lingwei Ma or Dawei Zhang.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025