

Bridging the Semantic-Numerical Gap: A Numerical Reasoning Method of Cross-modal Knowledge Graph for Material Property Prediction

Guangxuan Song^a, Dongmei Fu^{a,✉}, Zhongwei Qiu^{c,d}, Zijiang Yang^a,
Jiixin Dai^a, Lingwei Ma^b, and Dawei Zhang^{b,✉}

^aBeijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

^bNational Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing, China

^cAlibaba DAMO Academy, Hangzhou, China

^dZhejiang University, Hangzhou, China

Abstract

Using machine learning (ML) techniques to predict material properties is a crucial research topic. These properties depend on numerical data and semantic factors. Due to the limitations of small-sample datasets, existing methods typically adopt ML algorithms to regress numerical properties or transfer other pre-trained knowledge graphs (KGs) to the material. However, these methods cannot simultaneously handle semantic and numerical information. In this paper, we propose a numerical reasoning method for material KGs (NR-KG), which constructs a cross-modal KG using semantic nodes and numerical proxy nodes. It captures both types of information by projecting KG into a canonical KG and utilizes a graph neural network to predict material properties. In this process, a novel projection prediction loss is proposed to extract semantic features from numerical information. NR-KG facilitates end-to-end processing of cross-modal data, mining relationships and cross-modal information in small-sample datasets, and fully utilizes valuable experimental data to enhance material prediction. We further propose two new High-Entropy Alloys (HEA) property datasets with semantic descriptions. NR-KG outperforms state-of-the-art (SOTA) methods, achieving relative improvements of 25.9% and 16.1% on two material datasets. Besides, NR-KG surpasses SOTA methods on two public physical chemistry molecular datasets, showing improvements of 22.2% and 54.3%, highlighting its potential application and generalizability. We hope the proposed datasets, algorithms, and pre-trained models* can facilitate the communities of KG and AI for materials.

Keywords: knowledge graph, cross-modal learning, small-sample learning, graph neural network, material property prediction

1. Introduction

With the rise of big data in scientific and material research, the fourth paradigm of material data-driven research [1] is attracting widespread attention. In this context, machine learning (ML), with its outstanding modeling capabilities and reduced experimental costs, enables immense opportunities for predicting [2, 3] and exploring [4, 5] new materials based on their properties. However, ML in material science faces two challenges despite its achievements. First, high-quality material data is scarce due to the heavy reliance on experiments to obtain material properties. Material preparation and testing require substantial time and resources, leading to small data volume [6]. Second, effectively representing and utilizing multi-modal data is difficult [7]. Cross-modal ML requires a large amount of data, which is hard to obtain in material science. This dilemma often leads to the neglect of semantic information, such as processing techniques for material property prediction.

Two main approaches address the material property prediction challenges arising from the complexities of small-sample data and multi-modal data representation and utilization. Fig. 1 (a) shows the numeric modality ML methods comprising classical ML methods [8, 9] (e.g. random forests, RF) and neural network methods [10], based on material mechanisms and data

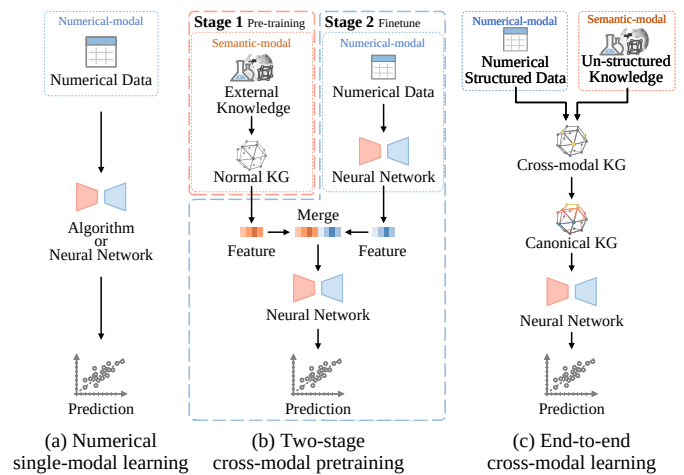


Figure 1: Comparison of property prediction methods. The methods in (a) ignore material semantic information. The methods in (b) lack integration of numerical and semantic aspects. Our method in (c) utilizes numerical and semantic information while accounting for inter-sample relationships in small sample data.

features. These methods aim to uncover relationships between material numerical features and properties. They have advan-

tages in small-sample problems but often overlook semantic information. Fig. 1 (b) illustrates cross-modal pre-training methods that incorporate supplementary information, including external knowledge [11, 12]. Pre-training with knowledge graphs (KGs) enhances ML by adding semantic knowledge, enabling better insights from small-sample data. KGs [13] efficiently model semantic information and organize knowledge. However, constructing efficient domain-specific KGs demands extensive data and research efforts. Furthermore, due to the limitations of two-stage methods, the absence of numerical data in the initial pre-training stage and the sole extraction of numerical features in the second numerical modeling stage limit the interaction between numerical and semantic information in feature extraction, resulting in a semantic-numerical gap. The reason KGs are often used to handle semantic data alone is that normal KG construction primarily depends on text, and KG reasoning is primarily centered around modeling the KG’s graph structure. Although certain research [14, 15] has focused on KG representation learning involving numerical attributes, their modeling of numerical attributes aims to enhance node or relationship representations, effectively aligning numerical values with the semantic space. While this enhances tasks such as link prediction, it still lacks the ability to effectively capture relationships between numerical values or the interactions between numerical and semantic information. In scientific domains that require quantitative descriptions, especially in material property prediction problems, conducting numerical reasoning on cross-modal KGs becomes a focal point of our attention. Simultaneously, both of the mentioned approaches in Fig. (a) and (b) suit small-sample material property prediction data. However, they exclusively focus on the relationship between material features and property, overlooking the relationship between samples in small-sample datasets.

To address the above problems, we propose the **Numerical Reasoning method of Knowledge Graph (NR-KG)**, as shown in Fig. 1 (c). Inspired by KG pre-training and the incorporation of numerical attributes into KG representation learning, we aim to introduce numerical information into KG construction for forming a cross-modal KG. Initially, NR-KG extracts numerical and semantic information from material data and creates a cross-modal KG. Subsequently, we propose a novel cross-modal KG projection learning method, projecting the cross-modal KG into a canonical KG. To bridge the semantic-numerical gap in cross-modal projection, NR-KG uses a comparative learning loss for unified cross-modal learning and proposes an innovative projection prediction loss to supervise the learning of projection features in cross-modal information. Ultimately, using graph neural networks (GNN), NR-KG propagates node information on the canonical KG based on real associations between materials, inferring material property from the material proxy node feature. The cross-modal KG structures diverse data with different processing or experimental conditions. Utilizing the projection learning of canonical KG, we fuse and model the semantic-numerical cross-modal information, establishing inter-sample relationships. NR-KG resolves the mentioned cross-modal representation and overlooked inter-sample relationship concerns. By comprehensively utilizing data, it en-

hances property prediction in small-sample scenarios without requiring extensive external knowledge pre-training.

Additionally, we propose two high-entropy alloys (HEA) datasets. HEAs are acclaimed in the scientific community for their exceptional mechanical properties, corrosion resistance, etc [16]. These datasets will facilitate research on ML techniques and new materials.

NR-KG is assessed using the HEA datasets. Moreover, NR-KG exhibits impressive performance in predicting molecular properties within the public dataset for scientific domain physical chemistry, underscoring its considerable potential in scientific data.

NR-KG is the first end-to-end KG numerical reasoning method, providing new insight into processing valuable experimental data for small-sample scientific data. Our contributions can be summarized as follows:

- We propose two HEA datasets on hardness and corrosion resistance, integrating existing public HEA datasets and adding HEA data and semantic information from literature, offering diverse data.
- Our novel NR-KG method merges numerical and semantic cross-modal data, pioneering an end-to-end KG approach for semantic-numerical data mining.
- We propose an innovative projection prediction loss using high-dimensional generalized F-point to address the semantic loss issue in numerical information encoding, effectively capturing numerical-semantic information.
- Extensive experiments show that NR-KG achieves state-of-the-art results and demonstrates excellent application scalability on two datasets related to High Entropy Alloys (HEA) in materials science and two publicly available molecular property datasets in the scientific domain. Ablation studies emphasize the effectiveness of each model component.

2. Related Work

In addressing the material property prediction problem, we explored cross-modal knowledge graphs to enhance prediction accuracy in small-sample scenarios. Here, we provide a brief overview of related works on these issues.

2.1. Material Property Prediction

The issue of utilizing machine learning (ML) for predicting material properties has been studied for decades. Due to the scarcity of material data and the high-dimensional complexity of material property spaces, researchers have focused on modeling problems with small-sample material data. Classical ML methods often outperform more complex deep learning networks on small-sample materials property prediction [17], due to their better generalization performance. Expert prior knowledge enhances ML predictive prowess by formulating material descriptors with authentic physical causal relationships [18],

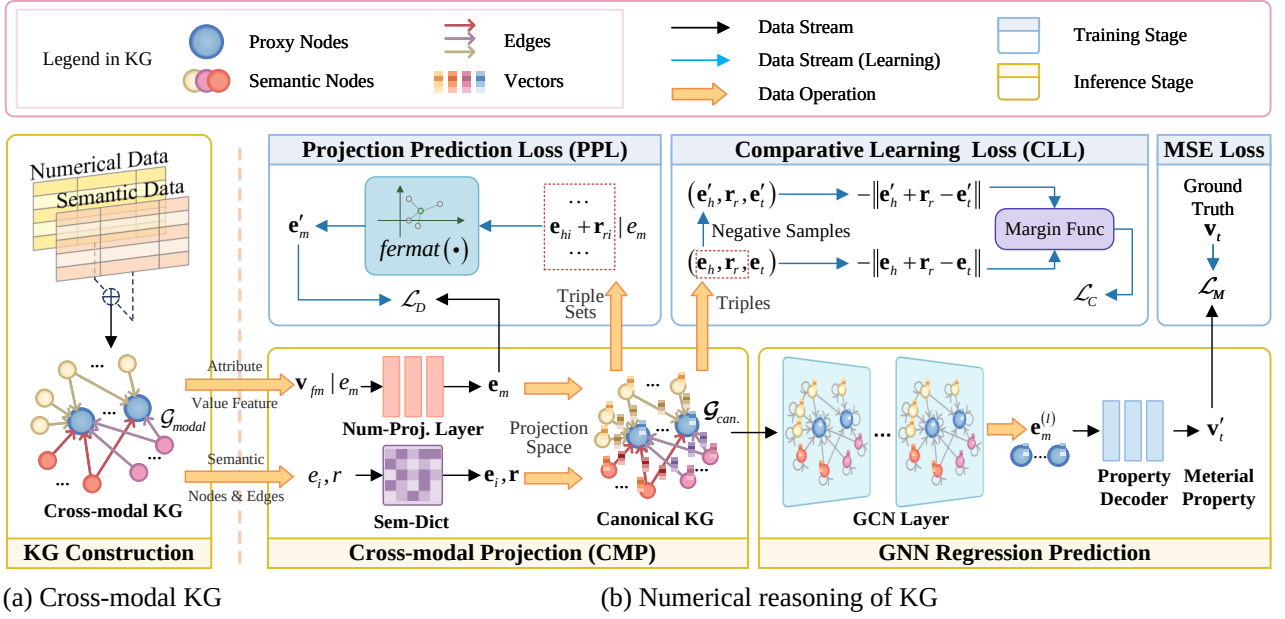


Figure 2: The NR-KG framework includes (a) KG construction and (b) numerical reasoning process. First, cross-modal KG $\mathcal{G}_m$ is constructed with numerical and semantic data. Then, proxy nodes represented by material numerical features $\mathbf{v}_{fm}|e_m$, semantic nodes e_i , and edges r in $\mathcal{G}_m$ are projected into canonical KG $\mathcal{G}_c$ using CMP. $\mathcal{G}_c$ is supervised by CLL and PPL. GNN Regression Prediction facilitates information exchange and predicts material properties $\mathbf{v}'$. $\oplus$, Num-Proj. Layer, Sem-Dict, Margin Func, and $\text{fermat}(\cdot)$ refer to the join operation, Numerical Projection Layer, Semantic Dictionary, Margin function, and high-dimensional generalized F-point calculation.

yielding valuable insights for small-sample ML. [19] introduces elemental information as external knowledge to enhance model predictive capabilities. While more efficient deep learning feature extractors, like Graph Neural Networks (GNN) [20, 21] or transformers [22], have been applied to material property prediction tasks, capturing complex features, they come with increased data requirements. However, such methods encounter difficulties in integrating semantic information into modeling.

2.2. KG Representation and Cross-modal Learning

KG representation learning captures low-dimensional representations of entities and relationships with their semantics [23, 24] and has certain extrapolation abilities [25]. Traditional KG representation learning methods, such as TransE [26] and other translation-based models, excel in representing semantic information, but their ability to represent numerical properties is limited. These methods can flexibly encode semantic information, showing potential applications in scientific fields such as medicine [27]. Recent research [28, 29, 14] tries to model numerical attributes in KGs to enhance semantic representation and improve link prediction accuracy. However, these methods mainly align numerical attributes (e.g. city latitude and longitude) embedding with the semantic embedding space, and do not address the problem of mining complex relationships between numerical and semantic values. End-to-end KG representation learning for joint reasoning of numerical and semantic information in scientific domains is still relatively limited.

2.3. Small-sample Learning

In materials science research, scenarios with limited sample data are extremely common. The challenge lies in capturing the crucial features of the data within these limited samples to avoid overfitting during training. Transfer learning, involving the transfer of knowledge from one domain or task (source) to another (target) [30], reduces the data requirements for the target domain and has mature applications. These methods can provide effective prior knowledge but necessitate a reasonably close distribution match between pre-training and target datasets. Additionally, during the pre-training phase, appropriate tasks need to be designed to capture features that generalize well for downstream tasks. Recent work has employed transfer learning strategies to introduce external knowledge and enhance predictive performance. In the [11, 12], knowledge was pre-trained on an ElementKG, and then this pre-trained external knowledge was incorporated into the process of predicting molecular properties. Nevertheless, these methods demand substantial data and computational resources, and the two-stage process creates a numerical-semantic gap, not fully leveraging the limited data. Furthermore, the potential value of inter-sample correlation information remains underexplored.

3. Method

To address the challenges of modeling semantic information and bridging the numerical-semantic gap in scientific small-sample data scenarios, as well as establishing relationships among samples and making optimal use of limited data, we

have defined the tasks of material property prediction, cross-modal KG numerical reasoning and numerical reasoning projection learning in Sec. 3.1. The NR-KG framework is proposed in Sec. 3.2, with technical details presented in Sec. 3.3, 3.4, and 3.5. Sec. 3.6 provides the complete algorithm for NR-KG.

3.1. Problem Formulation

3.1.1. Material Property Prediction

Material property prediction forecasts the physical and chemical characteristics of materials through a regression model that uses known features and their relationships. This process is mathematically represented as

$$v = F_{\text{property}}(X_p, \mathcal{R}_p), \quad (1)$$

where $v \in \mathbb{R}$ is the predicted material property value, X_p includes physical and chemical features, among other related information, $\mathcal{R}_p$ encompasses the relationships between these features, and the function F_{property} is the material property prediction model.

Unlike common sense KG numerical predictions, material property predictions are constrained by fundamental physical laws of the scientific world. The data-driven prediction process aims to fit complex physical models using data. In contrast, the numerical information in common sense KG numerical attribute (like latitude, longitude, year, etc.) modeling represents more semantic information, serving as context distinct from discrete semantic information (nodes or edges). Therefore, when considering the integration of semantic information, the numerical and semantic information in material property prediction differs in form and content across modalities.

3.1.2. Numerical Reasoning on Cross-modal KGs

KG is a knowledge model that structurally describes knowledge using nodes and edges. Let there be a node set $\mathcal{E}$, an edge set $\mathcal{R}$, an attribute set $\mathcal{A}$, and a real number set $\mathcal{V}$. Then, the cross-modal KG ($\mathcal{G}_m$) is represented as a set of triples

$$\mathcal{G}_m = \{(e_1, r, e_2)\} \cup \{(e, a, v)\}, \quad (2)$$

where $(e_1, r, e_2) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, $(e, a, v) \in \mathcal{E} \times \mathcal{A} \times \mathcal{V}$. In the attribute set $\mathcal{A}$, we define $\mathcal{A}_f \cup \mathcal{A}_t = \mathcal{A}$, where the subset $\mathcal{A}_f$ represents the feature attribute set that describes the features of relevant entities, while $\mathcal{A}_t$ represents the target attribute set that describes the properties of entities.

Numerical reasoning constructs two function mappings, $\psi : \mathcal{V} \rightarrow \mathcal{E}$ and $\Omega : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathcal{V}$, which predict the target attributes of entities based on their feature attributes and triple information.

3.1.3. Numerical Reasoning Projection Learning

Let us define the projection dimension as H . We define the function $f : \mathcal{V} \rightarrow \mathbb{R}^H$ to map the node's feature attribute $\mathbf{v}$ to its canonical vector $f(\mathbf{v})$, and the function $g : \mathbb{R}^H \times \mathbb{R}^H \times \mathbb{R}^H \rightarrow \mathcal{V}$ to map the triple information $(\mathbf{e}_1, \mathbf{r}, \mathbf{e}_2)$ to a vector relevant to the target attribute value. Thus, numerical reasoning projection learning on $\mathcal{G}_m$ is defined as $\psi(v) \xrightarrow{\text{def}} f(\mathbf{v})$ and $\Omega(e_1, r, e_2) \xrightarrow{\text{def}} g(\mathbf{e}_1, \mathbf{r}, \mathbf{e}_2)$.

3.2. Framework

The structure of **NR-KG** is shown in Fig. 2. A cross-modal KG $\mathcal{G}_m$ is constructed through the Cross-modal KG Construction process. Then, $\mathcal{G}_m$ is projected into the canonical KG, denoted as $\mathcal{G}_c$, by the Cross-modal Project (CMP). The supervision involves two primary components: Projection Prediction Loss (PPL) and Comparative Learning Loss (CLL). Lastly, GNN Regression Prediction is utilized for the forecast of properties represented by material proxy nodes. Notably, our framework is designed as an end-to-end process, fostering the integration of numerical and semantic data within the ($\mathcal{G}_m$). NR-KG approach ensures that the model captures intricate relationships and patterns in the data, leading to more accurate predictions and a deeper understanding of material properties.

3.3. Cross-modal KG Construction

We classify material data into two types: numerical data (e.g., material composition) and semantic data. The latter encompasses text-based descriptions of material manufacturing processes, structures, or classification information from various perspectives. Given the flexible nature of semantic data, which is influenced by recorder subjectivity and the presence of multiple process combinations or missing information, a cross-modal KG is proposed to capture the versatile relationships between numerical and semantic data in material datasets.

The process of cross-modal KG ($\mathcal{G}_m$) construction is outlined as follows: To incorporate numerical data, we set a proxy node e_m for each material. Numerical attributes characterizing material properties are subsequently added as node feature attributes, denoted as a triple set $\{(e_m, r_f, v_{fm})\}$, where r_f and v_{fm} denote the feature attribute and the numerical feature value of the material, respectively. These numerical feature value represent the independent variables that exert influence on the properties of the material. Moreover, numerical data describing material properties is added as target attribute values for e_m , denoted as a triple set $\{(e_m, r_t, v_{tm})\}$, where r_t and v_{tm} correspond to the target attribute and numerical target attribute value of e_m . These values correspond to the dependent variables interlinked with the properties of the material.

Concerning semantic information, we extract named entities and treat them as separate nodes. These nodes are then linked to relevant material proxy nodes, creating semantic associations denoted as (e_m, r_k, e_i) or (e_i, r_l, e_j) , where e_i and e_j signify different nodes, and r_k and r_l represent different edges within $\mathcal{G}_m$. Additionally, We perform an operation

$$(e_i, r_k, e_m) \leftarrow (e_m, r_k, e_i), \quad (3)$$

to reverse relationships originating from material nodes. This aligns with the GNN Regression Prediction in NR-KG, transforming relationship descriptions from the "active voice" to the "passive voice".

Cross-modal KG Construction introduces material proxy nodes as repositories of numerical data. This effectively organizes data in numerical-modal and semantic-modal forms, which lack an inherent structure, into a cohesive KG.

3.4. Cross-modal KG Projection to Canonical KG

Cross-modal Projection Module maps each node and edge in the cross-modal KG ($\mathcal{G}_m$) to a projection space $\mathbb{R}^H$, referred to as the canonical KG ($\mathcal{G}_c$), which plays a pivotal role in bridging the semantic-numerical gap. The projection dimension is denoted by H . For the semantic nodes e_i and edges r , they are independently mapped to vectors $\mathbf{e}_i \in \mathbb{R}^H$ and $\mathbf{r} \in \mathbb{R}^H$ by a semantic dictionary:

$$\begin{aligned} \mathbf{e}_i &= \text{Norm}(f_{\text{SemD}}(e_i; E_s)) \\ \mathbf{r} &= f_{\text{SemD}}(r; E_s) \end{aligned}, \quad (4)$$

where E_s denotes the learnable parameters, function f_{SemD} represents the mapping of the semantic dictionary, and Norm represents the normalization operation to ensure $\mathbf{e}_i$ is within a reasonable range, stabilizing model training. For the proxy nodes e_m , we propose the numerical projection layer to predict their vectors $\mathbf{e}_m$ in $\mathcal{G}_c$ based on the vector of node feature attributes $\mathbf{v}_{fm}$. This ensures that the projection vector of e_m is independent of the semantic dictionary. The implementation of the numerical projection layer employs the multi-layer perceptron (MLP) network in NR-KG, as MLP can preserve numerical features:

$$\mathbf{e}_m = \text{Norm}(f_{\text{NPL}}(\mathbf{v}_{fm}; \theta_{\text{NPL}})), \quad (5)$$

where f_{NPL} is the implementation of the numerical projection layer, and θ_{NPL} is network parameters.

To ensure that $\mathcal{G}_c$ after cross-modal projection can simultaneously fuse numerical and semantic information, we were inspired by the translation-based method TransE [26] to design the comparison learning loss (CLL). This loss assumes that $\mathbf{e}_h$ can be translated from $\mathbf{e}_t$ based on $\mathbf{r}_r$ in the triple (e_h, r_r, e_t) . For the triples (e_m, r_k, e_i) or (e_i, r_l, e_j) in $\mathcal{G}_m$, we uniformly refer to them as (e_h, r_r, e_t) . The CLL defines a distance function $d(e_h, r_r, e_t) = -\|\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t\|$ to evaluate the plausibility of the triple (e_h, r_r, e_t) , where $\|\cdot\|$ represents the L2-norm. The goal of comparative learning within $\mathcal{G}_c$ is to minimize a margin loss, given by:

$$\mathcal{L}_C = \begin{cases} d(e_h, r_r, e_t) & , (e_h, r_r, e_t) \in \mathcal{G}_m \\ \max(0, \gamma - d(e'_h, r_r, e'_t)) & , (e'_h, r_r, e'_t) \notin \mathcal{G}_m \end{cases}, \quad (6)$$

where $\gamma > 0$ serves as a hyperparameter that differentiates between positive and negative samples, (e'_h, r_r, e'_t) denotes negative samples extracted from $\mathcal{G}_m$. A feasible method for constructing these negative samples involves randomly substituting the e_h and e_t in the positive triples (e_h, r_r, e_t) .

Predicting the projection vector $\mathbf{e}_m$ by learning the numerical data of the proxy node e_m remains a challenge, primarily due to the common situation where the projection dimension H is much greater than the dimension of $\mathbf{v}_{fm}$. Additionally, there's a lack of semantic information during the process of predicting $\mathbf{e}_m$ from $\mathbf{v}_{fm}$ using the numerical projection layer. Hence, we propose a reverse method that involves directly computing the projection vector of e_m using the concept of CLL. This method is termed projection prediction loss (PPL) and serves to enhance the supervision of the numerical projection layer in learning $\mathcal{G}_c$ predictions.

PPL is built on a fundamental assumption from CLL: $\mathbf{e}_h + \mathbf{r}_r \approx \mathbf{e}_t$. As we place e_m in the tail entity slot while constructing $\mathcal{G}_m$, the computed value for the specific $\mathbf{e}_m$ in the $\mathcal{G}_c$ is defined as an optimization problem:

$$\mathbf{e}'_m = \arg \min_{\mathbf{e} \in \mathbb{R}^H} \sum_{\mathbf{p} \in \mathcal{P}} \|\mathbf{e} - \mathbf{p}\|, \quad (7)$$

where $\mathcal{P}$ means all sets of $\mathbf{e}_h + \mathbf{r}_r$ in $\{(\mathbf{e}_h, \mathbf{r}_r, \mathbf{e}_t)\}$ for a specific $\mathbf{e}_t$. Thus, our objective is to find a point among all points $\mathbf{e}'_m$ that minimizes the sum of distances of all points within $\mathcal{P}$. This point $\mathbf{e}'_m$ corresponds to a high-dimensional generalized F-point [31], which can be solved through optimization methods. Due to the geometric complexity of high-dimensional space, the high-dimensional generalized F-point typically lacks an explicit solution, and numerical optimization methods are used to obtain its solution within an acceptable range. In NR-KG, for computational efficiency, we employed the average point of $\mathbf{p}$ to estimate $\mathbf{e}'_m$.

To learn the projection vector $\mathbf{e}_m$ of proxy nodes, PPL minimizes a distance loss given by:

$$\mathcal{L}_D = \sum_{\mathbf{e}_m \in \mathcal{E}} \|\mathbf{e}_m - \mathbf{e}'_m\|. \quad (8)$$

The losses $\mathcal{L}_C$ and $\mathcal{L}_D$ govern the process of cross-modal KG projection to the canonical KG, establishing connections between numerical and semantic information, ensuring alignment within the canonical KG. In the learning process of the canonical KG, the projection vector $\mathbf{e}_m$ for proxy nodes e_m is defined as an encoding process that depends solely on numerical features. This process, compared to traditional KG embedding methods (e.g., [26]), allows flexible incorporation of new proxy nodes during the inference stage, which holds practical significance for applications in the materials domain.

3.5. GNN Regression Prediction

The target features of proxy nodes depend not only on the proxy nodes themselves but also on their neighboring nodes. To model this relationship, we utilize the graph convolutional network (GCN) to propagate node feature vectors in accordance with real-world information relationship logic. During this process, we disregard the relations vectors in canonical KG and treat them as connections. The GCN operation is as follows:

$$\mathbf{e}^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \mathbf{e}^{(l)} \mathbf{W}), \quad (9)$$

where, $\tilde{A} = A + I$, which A is the critical matrix, $\tilde{A}$ represents the addition of self-loops, and $\tilde{D}$ stands for the degree matrix of $\tilde{A}$, and $\sigma(\cdot)$ denotes the activation function.

Finally, we use an MLP network as a property decoder to predict the target attribute values $\mathbf{v}_t$ based on the proxy node features:

$$\mathbf{v}'_t = f_{PD}(\mathbf{e}_m; \theta_{PD}), \quad (10)$$

where function f_{PD} is the property decoder, θ_{PD} is network parameters, and $\mathbf{v}'_t$ denotes the prediction of the target attribute values $\mathbf{v}_t$.

This step involves learning a mean squared error (MSE) loss, defined as

$$\mathcal{L}_M = \sum_{v_t \in V} \|\mathbf{v}'_t - \mathbf{v}_t\|, \quad (11)$$

where $\|\cdot\|$ denotes the L2-norm.

Algorithm 1 NR-KG Algorithm

Input D : dataset of material information.

Parameter $\gamma, \gamma_a, \gamma_b$: the hyperparameter in $\mathcal{L}$. H : canonical KG ($\mathcal{G}_c$) dimension. T : maximum iteration.

Output $\mathbf{v}'_t$: predicted target attribute value for proxy node.

Training stage:

- 1: Construct a cross-modal KG $\mathcal{G}_m$ from D .
- 2: Initialize the semantic dictionary E_s
- 3: **for** $epoch = 1$ to T **do**
- 4: **for** $(e_h, r_r, e_t) \in \mathcal{G}_m$ **do**
- 5: **for** $t \in (e_h, r_r, e_t)$ **do**
- 6: **if** $t \in \{e_m\}$ **then** $\mathbf{t} \leftarrow \text{Norm}(f_{NPL}(\mathbf{v}_{fm}))$
- 7: **else** $\mathbf{t} \leftarrow \text{Norm}(f_{semD}(t, E_s))$ or $f_{semD}(t, E_s)$
- 8: **end for**
- 9: **end for**
- 10: $\mathcal{L}_C \leftarrow \sum \max(0, d(e'_h, r_r, e'_t) - d(e_h, r_r, e_t) + \gamma)$
- 11: $\mathcal{L}_D \leftarrow \sum \|\mathbf{e}_m - \arg \min \sum \|\mathbf{e} - \mathbf{p}\|\|$
- 12: $\mathcal{G}_c \leftarrow f_{GCN}(G)$ {Ref Eq.(9)}
- 13: $\mathbf{v}'_{fm} \leftarrow f_{PD}(\mathbf{e}_m)$ **for** each e_m in $\mathcal{G}_c$
- 14: $\mathcal{L}_M \leftarrow \sum \|\mathbf{v}'_t - \mathbf{v}_t\|$
- 15: Update the model w.r.t. $\mathcal{L} = \mathcal{L}_M + \gamma_a \mathcal{L}_C + \gamma_b \mathcal{L}_D$
- 16: **end for**
- 17: **return** $\mathbf{v}'_t$

Inference stage:

- 1: $\mathcal{G}_m \xleftarrow{\text{add}} \{(e_m, r_f, \mathbf{v}_{fm})\}$
 - 2: Same as lines 4 to 9, 12, 13 and 17 in the training stage
 - 3: **return** $\mathbf{v}'_t$
-

3.6. Learning

3.6.1. Algrithom

The training and inference stages are shown in Algorithm 1. In the training stage, during each iteration, the cross-modal KG is initially projected onto the canonical KG, followed by GCN calculations to determine the material property values of the material proxy nodes. Subsequently, the loss is calculated to update the network parameters. In the testing stage, after adding the predicted material entities and relevant relationships to the cross-modal KG, the material property can be obtained using the trained network.

3.6.2. Loss Function

The overall loss of NR-KG is denoted as

$$\mathcal{L} = \mathcal{L}_M + \gamma_a \mathcal{L}_C + \gamma_b \mathcal{L}_D, \quad (12)$$

where $\mathcal{L}_M$ denotes the mean squared error loss, $\mathcal{L}_C$ denotes the comparative learning loss, and $\mathcal{L}_D$ denotes the projection prediction loss, γ_a and γ_b are coefficients of $\mathcal{L}_C$ and $\mathcal{L}_D$.

4. Experiments

To validate the performance, effectiveness, and generalizability of NR-KG, comprehensive experiments were conducted. Sec. 4.1 details the experimental settings, focusing on the two new HEA datasets in Sec. 4.1.1. Sec. 4.2 describes the construction process and statistical information of the cross-modal KG. Comparative experiments between NR-KG and state-of-the-art (SOTA) methods on our HEA datasets, highlighting NR-KG’s outstanding performance, are presented in Sec. 4.3. 4.4 and 4.5 respectively provide insights into ablation and interpretability experiments, elucidating NR-KG’s working principles. To demonstrate NR-KG’s application scalability, Sec. 4.6 presents comparative experiments on two publicly available molecular datasets.

4.1. Experiments Settings

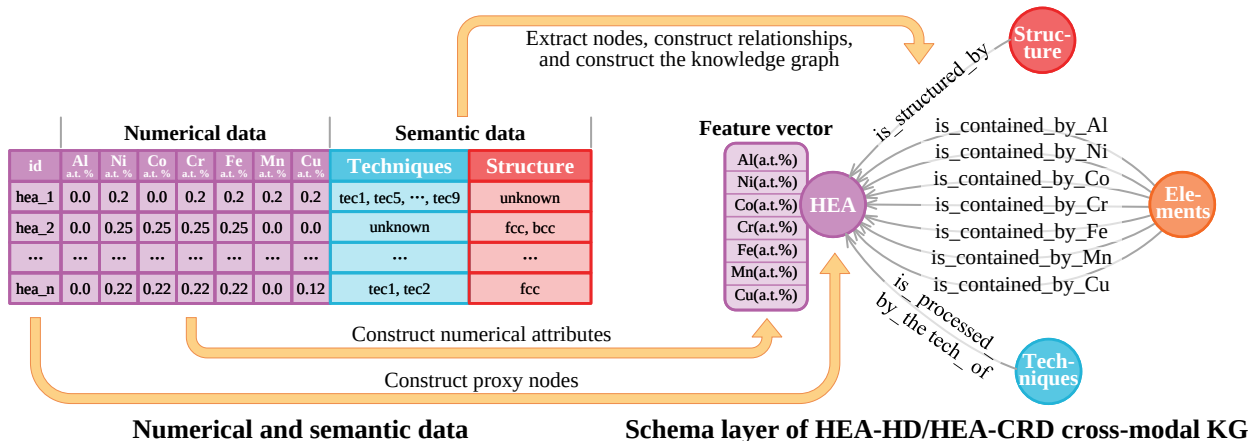
4.1.1. HEA Dataset Construction and Dataset Selection

The experimental dataset comprises two categories: two high-entropy alloy (HEA) datasets constructed by us to demonstrate the effectiveness of NR-KG in material property prediction tasks, and two publicly available physical chemistry molecule property datasets used to validate the application scalability of NR-KG.

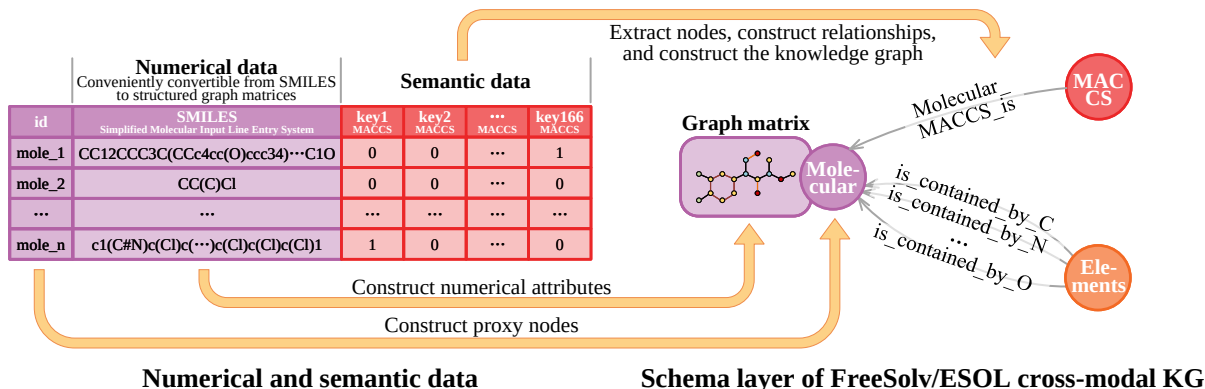
HEA datasets comprise HEA public datasets and newly collected data. We focused on HEAs with Al-Fe-Ni-Co-Cr-Cu-Mn elements due to extensive research in materials science. We combined data from [32, 33, 34], adding semantic information (processing techniques and crystal structures) from the original literature. Additionally, we enriched the dataset with recent high-level publications in materials research from Scopus. During extraction, GPT-3.5 [35] aided in locating and extracting key information, verified and added by materials experts. Original expressions were retained without vocabulary alignment to assess method effectiveness in the presence of semantic noise (e.g. errors, omissions, ambiguities).

We propose the HEA hardness dataset (HEA-HD) and the HEA corrosion resistance dataset (HEA-CRD). All materials include elemental atomic ratios, while processing techniques and crystal structure information have some missing values. HEA-HD comprises 397 items labeled with HV (higher values indicating greater hardness). HEA-CRD contains 151 items labeled with logarithmically processed corrosion current $\ln(I_{corr})$ (lower values indicating better corrosion resistance). Details of data collection, cleaning, and a detailed dataset analysis are provided in Appendix A.2.1.

Due to limitations in previous algorithms in handling semantic information, most publicly available material datasets lack semantic data aligned with numerical information. To validate the applicability of NR-KG, we conduct experiments on a public benchmark dataset FreeSolv [36] and ESOL [37]. Prediction of molecular properties is a central research topic in physics, chemistry, and materials science that focuses on predicting various properties and characteristics of molecules based on their structural representations. Therefore, the task of predicting molecular properties is used to evaluate the applicability of NR-KG in the field of science.



(a) KG Construction for HEA Composition Data



(b) KG Construction for Molecular Data

Figure 3: Cross-modal KG construction and schema layer for HEA and molecular data. (a) and (b) show the process of HEA and molecular data, respectively. In the tables of numerical and semantic data and in the schema layer of the cross-modal KG, the same colors indicate identical categories of information.

4.1.2. Data Split and Evaluation Metrics

For HEA-HD and HEA-CRD datasets, we adopted the 6-fold cross-validation method: The dataset was split into 6 equal parts, cyclically using 4 parts for training and the remaining 2 parts for validation and testing. During splitting, we ensured the training set included all semantic words to construct a complete semantic dictionary.

For FreeSolv and ESOL datasets, we apply the scaffold splitting [38] method. Scaffold splitting splits the molecules with distinct two-dimensional structural frameworks into different subsets, providing a more challenging but practical setting since the test molecule can be structurally different from the training set. Following [39, 40, 11], the data were split into training, validation, and testing sets at an 8:1:1 ratio using scaffold splitting. This random partitioning was repeated three times to ensure the statistical validity of the experimental results.

Two types of evaluation metrics were employed:

- For HEA property prediction experiments (numerical regression task): mean squared error (MSE, smaller is better), mean absolute error (MAE, smaller is better), and coefficient of determination (R^2 , larger is better). Additionally,

for comparability with public results, root mean square error (RMSE, smaller is better) is used to assess molecular property prediction performance.

- For model interpretability experiments (link prediction task): mean reciprocal rank (MRR, larger is better), mean rank (MR, lower is better), and Hits@1 (top one prediction hit rate, larger is better).

4.2. Cross-modal KG Construction and Analysis

We developed scripts to transform the datasets into triples (a structured KG format) as shown in Fig. 3. For HEA data, HEA-HD and HEA-CRD were processed into triples where different processing techniques, crystal structures, and elements are considered as different nodes. Each HEA is assigned a HEA proxy node, linked to processing techniques, crystal structures, and element nodes through relations in “passive voice”. Element composition content become feature attribute values for proxy nodes, and property values are treated as target attribute values.

For molecular data, given the absence of direct semantic information in public datasets, we treated the inclusion relationships of molecular substructures as semantic data to test our

Table 1: Scale of the cross-modal KGs. “Unconnected HEA Nodes” and “Unconnected molecular nodes” represent the count of HEA nodes and molecular nodes that are not linked to any processing, crystal structure, or MACCS nodes, respectively.

(a) Cross-modal KGs of HEA datasets.			(b) Cross-modal KGs of molecular datasets.		
Items	HEA-HD	HEA-CRD	Items	FreeSolv	ESOL
HEA nodes	397	151	Molecular nodes	642	1128
Element nodes	7	7	Element nodes	9	9
Processing techniques nodes	33	51	MACCS nodes	147	147
Crystal structure nodes	7	9	Relationship category	10	10
Relationship category	9	9	Average degree of nodes	28.048	45.925
Average degree of nodes	7.191	6.046	Unconnected molecular nodes	0	0
Unconnected HEA nodes	194	15			

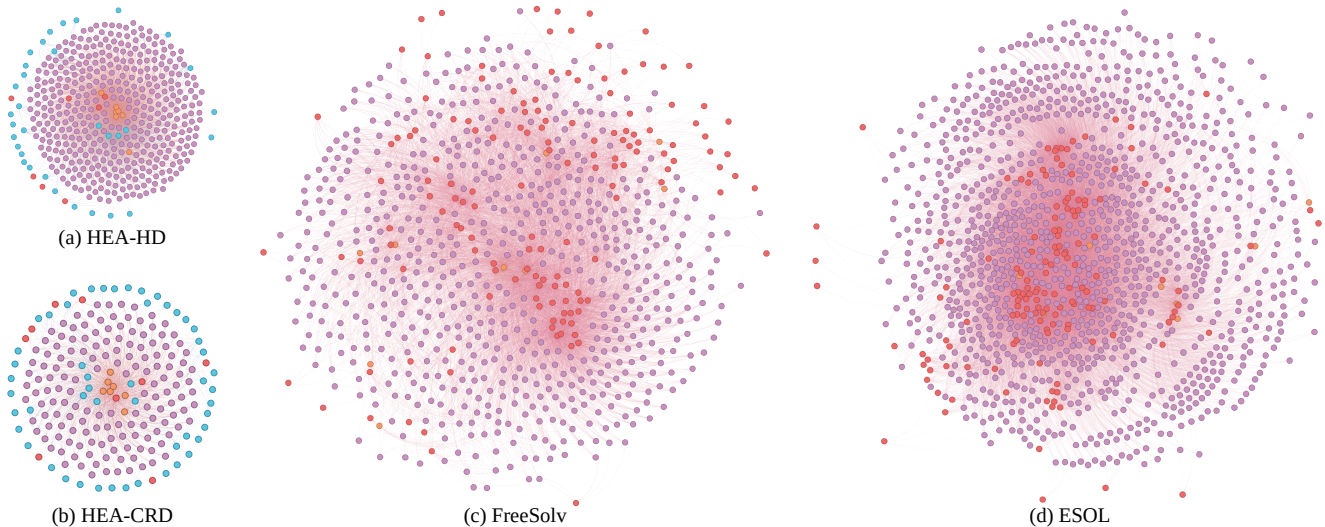


Figure 4: Visualization of cross-modal KGs. (a), (b), (c) and (d) depict the visualization of cross-modal KGs constructed from the HEA-HD, HEA-CRD, FreeSolv and ESOL datasets, respectively.

NR-KG method. Additionally, we considered the molecular structure matrices—including transition matrices and node attribute matrices—as numerical data. This step is crucial for creating molecular proxy nodes. Defined by the Molecular ACCess System (MACCS) keys [41] as semantic nodes, the 166 substructures facilitated connections between these molecular proxy nodes and semantic nodes based on their molecular MACCS fingerprints [42]. This method quickly created semantic links between molecules and utilized the RDKit open-source toolkit [43] for comprehensive molecular structure analysis.

Table 1 and Fig. 4 respectively show the statistical scale and visualization of cross-modal KGs. The HEA dataset is derived from real experiments in materials science publications. Consequently, some HEA nodes link to multiple processing techniques and crystal structure nodes, while others remain isolated. These cross-modal KGs effectively capture these complex relationships.

In the NR-KG training stage, HEA nodes in the validation and testing sets are temporarily masked to prevent their influence, forming the training cross-modal KG. These nodes are reintroduced during the validation and testing phases. However, some KG-based methods in comparison experiments cannot introduce new nodes during testing. For these methods, we only mask the target property of the HEA in the valida-

tion and testing sets, without deleting any nodes. This approach allows these methods to utilize semantic data from the validation and testing sets during training, thus reducing the difficulty of performance prediction tasks. Consequently, NR-KG tasks more closely reflect real-world scientific data prediction scenarios and pose greater challenges.

4.3. Comparison Experiments

4.3.1. Baselines

We evaluate the performance of NR-KG and compare it with the following four categories of methods:

- Classical ML-based Methods (CMMs): YSRF [44], HDSVR [45], and PCHMLP [46, 47]. We also verified the effectiveness of these methods when using one-hot encoding to model semantic information.
- Incorporating Element Knowledge Method (IEKM): DS-HEA [19].
- KG Numerical Attribute Modeling Methods (KG-NAMMs): KEN [28], MrAP [29], LiteralE [14], mkbe [48] and TransEA [15].
- Two-stage Pretrained KG Augmentation Method (KG-AM): KANO [11].

Table 2: Comparison experiment results. Results are shown as “mean $\pm$ standard deviation”. Bold cells highlight the best results per metric. The “New nodes” column signifies the method’s ability to predict new HEAs without retraining. CMM, IEKM, KG-NAMM, and KG-AM denote classical ML-based methods, incorporating element knowledge method, KG numerical attribute modeling methods, and KG augmentation method, respectively. Our method, NR-KG, achieved a 25.9% and 16.1% improvement over the second-best algorithm of MSE.

Datasets			HEA-HD			HEA-CRD		
Task			Regression (Testing set)			Regression (Testing set)		
Category	Methods	New nodes	MSE($\times 10^3$) $\downarrow$	MAE $\downarrow$	R ² $\uparrow$	MSE $\downarrow$	MAE $\downarrow$	R ² $\uparrow$
CMM	YSRF	✓	5.438 $\pm$ 1.706	46.853 $\pm$ 6.837	0.820 $\pm$ 0.031	3.895 $\pm$ 1.592	1.485 $\pm$ 0.350	0.202 $\pm$ 0.271
	YSRF(semantic)	✓	4.825 $\pm$ 1.604	43.932 $\pm$ 7.564	0.841 $\pm$ 0.036	3.015 $\pm$ 1.260	1.313 $\pm$ 0.246	0.362 $\pm$ 0.278
	HDSVR	✓	6.441 $\pm$ 1.312	63.555 $\pm$ 3.958	0.783 $\pm$ 0.039	5.039 $\pm$ 2.149	1.747 $\pm$ 0.281	-0.064 $\pm$ 0.362
	HDSVR(semantic)	✓	5.773 $\pm$ 1.238	61.066 $\pm$ 6.175	0.804 $\pm$ 0.045	2.816 $\pm$ 1.110	1.284 $\pm$ 0.152	0.408 $\pm$ 0.184
	PCHMLP	✓	6.017 $\pm$ 1.976	58.110 $\pm$ 9.069	0.801 $\pm$ 0.046	6.633 $\pm$ 1.437	2.125 $\pm$ 0.220	-0.426 $\pm$ 0.304
	PCHMLP(semantic)	✓	5.557 $\pm$ 1.483	55.417 $\pm$ 7.371	0.810 $\pm$ 0.057	2.633 $\pm$ 0.986	1.161 $\pm$ 0.146	0.449 $\pm$ 0.168
IEKM	DS-HEA	✓	6.632 $\pm$ 2.018	62.633 $\pm$ 13.743	0.770 $\pm$ 0.093	5.502 $\pm$ 0.853	1.809 $\pm$ 0.200	-0.096 $\pm$ 0.300
KG-NAMM	KEN	×	55.575 $\pm$ 9.557	187.591 $\pm$ 21.505	-0.826 $\pm$ 0.295	7.552 $\pm$ 0.702	2.229 $\pm$ 0.165	-0.675 $\pm$ 0.464
	MrAP	×	26.589 $\pm$ 1.930	141.571 $\pm$ 6.451	-0.016 $\pm$ 0.018	4.910 $\pm$ 0.840	1.941 $\pm$ 0.200	-0.473 $\pm$ 1.147
	LiteralE	×	23.783 $\pm$ 0.427	125.736 $\pm$ 4.110	0.167 $\pm$ 0.066	4.471 $\pm$ 1.068	1.744 $\pm$ 0.218	0.064 $\pm$ 0.281
	mkbe	×	28.314 $\pm$ 1.913	145.974 $\pm$ 7.222	-0.017 $\pm$ 0.053	4.572 $\pm$ 1.125	1.716 $\pm$ 0.316	-0.102 $\pm$ 0.099
	TransEA	×	33.386 $\pm$ 2.922	158.248 $\pm$ 6.959	-0.186 $\pm$ 0.098	5.270 $\pm$ 1.032	1.858 $\pm$ 0.262	-0.200 $\pm$ 0.173
KG-AM	KANO	✓	4.753 $\pm$ 1.070	50.605 $\pm$ 4.782	0.839 $\pm$ 0.032	2.746 $\pm$ 1.371	1.216 $\pm$ 0.193	0.420 $\pm$ 0.223
-	NR-KG(Ours)	✓	3.520$\pm$0.931	42.962$\pm$4.389	0.881$\pm$0.029	2.210$\pm$0.949	1.122$\pm$0.164	0.533$\pm$0.152

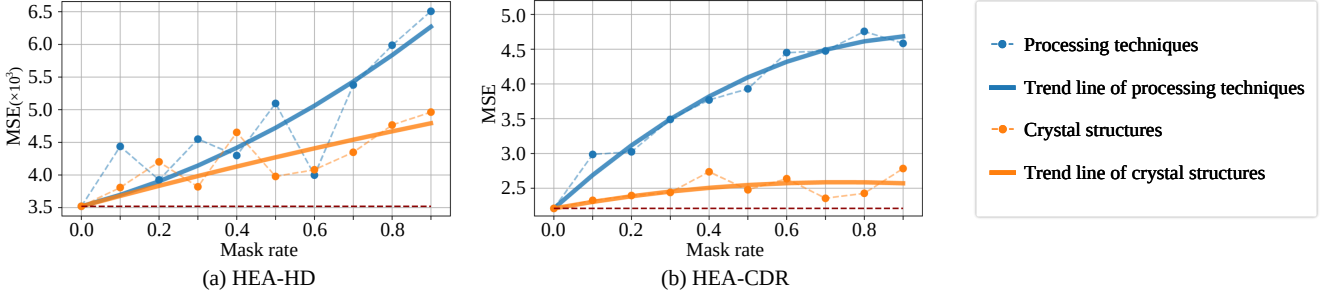


Figure 5: Semantic edges masking experimental results. (a) and (b) represent experimental results on HEA-HD and HEA-CRD datasets, respectively. As the masking ratio of semantic edges increases, the predictive MSE of NR-KG shows an upward trend, highlighting the significance of semantic information.

4.3.2. Experimental Results

As shown in Table 2, the results demonstrate that NR-KG achieved the best performance across all evaluation metrics. Specifically, NR-KG achieved a 25.9% and 16.1% improvement over the second-best method in terms of MSE on the HEA-HD and HEA-CRD, respectively.

The experimental results show that under the constraint of limited small-sample data, CMMs exhibit proficiency in capturing numerical features. Upon incorporating semantic information through one-hot encoding, regression accuracy improves, indicating a correlation between semantic information and material properties. IEKM introduces additional elemental knowledge but also significant redundancy in predicting HEA properties, considering only seven types of elements, and lacks sufficient constraint information from small-sample data. The numerical modeling of KG-NAMMs is semantic-oriented and exhibits poor modeling performance for material data with highly nonlinear features. Moreover, these methods mostly cannot introduce new nodes during the testing phase. To enable normal training and testing, although new nodes’ semantic information was included during the training of the KG-NAMMs method to reduce task difficulty, the performance in predicting

material properties is still very unsatisfactory. This indicates that merely aligning numerical values to the KG space cannot effectively handle the gap between numerical information and semantic information. While KG-AM shows potential, it integrates semantic information into the numerical prediction process through a pretrained KG. However, the two-stage method lacks interaction between numerical and semantic information, limiting its performance. As an end-to-end approach, **NR-KG** simultaneously mines numerical and semantic information, effectively leveraging data content, and achieves state-of-the-art performance.

4.4. Ablation Study

4.4.1. Integrity of Semantic Information

This experiment demonstrates NR-KG’s predictive accuracy under varying percentages of missing semantic nodes. During the training stage, we randomly masked a certain percentage of HEA nodes’ semantic edges related to processing techniques or crystal structures to simulate incomplete semantic information. Fig. 5 shows the trend of declining predictive accuracy for NR-KG on both HEA datasets as semantic information is absent. Fluctuations might stem from randomness introduced

Table 3: Experimental results on the effectiveness of model structure design. Results are shown as “mean $\pm$ standard deviation”. Bold cells highlight the best results per metric.

Datasets			HEA-HD			HEA-CRD		
Task			Regression (Testing set)			Regression (Testing set)		
Methods	Module	Training	MSE($\times 10^3$) $\downarrow$	MAE $\downarrow$	R ² $\uparrow$	MSE $\downarrow$	MAE $\downarrow$	R ² $\uparrow$
MLP	MLP base	end-to-end	6.017 $\pm$ 1.976	58.110 $\pm$ 9.069	0.801 $\pm$ 0.046	6.633 $\pm$ 1.437	2.125 $\pm$ 0.220	-0.426 $\pm$ 0.304
MLP	+semantic		5.557 $\pm$ 1.483	55.417 $\pm$ 7.371	0.810 $\pm$ 0.057	2.633 $\pm$ 0.986	1.161 $\pm$ 0.146	0.449 $\pm$ 0.168
normal-KG	KG base	two-stage	33.264 $\pm$ 3.963	148.058 $\pm$ 9.167	-0.129 $\pm$ 0.166	6.519 $\pm$ 1.028	2.086 $\pm$ 0.180	-0.435 $\pm$ 0.400
NR-KG-MLP	+numerical value		15.859 $\pm$ 7.708	103.799 $\pm$ 28.771	0.436 $\pm$ 0.343	4.612 $\pm$ 1.300	1.772 $\pm$ 0.284	0.039 $\pm$ 0.111
NR-KG-GCN	+GCN		10.636 $\pm$ 2.609	77.610 $\pm$ 13.947	0.638 $\pm$ 0.100	2.731 $\pm$ 1.207	1.224 $\pm$ 0.164	0.431 $\pm$ 0.180
normal-KG	KG base	end-to-end	32.926 $\pm$ 6.386	149.758 $\pm$ 16.358	-0.103 $\pm$ 0.091	6.145 $\pm$ 1.299	2.056 $\pm$ 0.219	-0.312 $\pm$ 0.226
NR-KG-MLP	+numerical value		4.535 $\pm$ 1.236	49.187 $\pm$ 7.529	0.849 $\pm$ 0.032	3.657 $\pm$ 1.269	1.529 $\pm$ 0.304	0.247 $\pm$ 0.176
NR-KG-GCN	+GCN		3.520$\pm$0.931	42.962$\pm$4.389	0.881$\pm$0.029	2.210$\pm$0.949	1.122$\pm$0.164	0.533$\pm$0.152

Table 4: Experimental results on effect of $\mathcal{L}_C$ and $\mathcal{L}_D$. Results are presented as the mean MSE from 6-fold cross-validation. Bold cells indicate the best results in each row.

Datasets	γ_a		0.00	0.04	0.08	0.12	0.16	0.20	0.30	0.40	0.60	0.80	1.00
HEA-HD	MSE($\times 10^3$) $\downarrow$	$\gamma_b=0.00$	4.108	4.038	3.853	3.794	3.620	3.780	3.787	3.840	3.908	3.902	3.944
HEA-CRD	MSE $\downarrow$		2.681	2.400	2.343	2.324	2.303	2.335	2.346	2.377	2.455	2.490	2.555
Datasets	γ_b		0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
HEA-HD	MSE($\times 10^3$) $\downarrow$	$\gamma_a=0.16$	3.620	3.567	3.520	3.728	4.081	4.189	4.173	4.170	4.204	4.430	4.471
HEA-CRD	MSE $\downarrow$		2.303	2.210	2.498	2.603	2.634	2.699	2.678	2.691	2.745	2.747	2.873

by the small dataset size. This observation underscores the significance of incorporating semantic information, such as processing techniques and crystal structures, for accurate material property prediction. The greater the information’s comprehensiveness, the more effectively NR-KG leverages it to enhance predictive performance.

4.4.2. Ablation Analysis of Components

We explored the impact of different components of the NR-KG approach. A baseline reference was established using MLP, a neural network-based numerical modeling method. For the “normal-KG” approach, a normal KG was constructed solely using semantic information, and TransE [26] was employed for representation learning. In the NR-KG-MLP variant, the GCN in NR-KG was replaced with MLP. The complete NR-KG method encompassing all components was also evaluated, referred to as NR-KG-GCN for differentiation. Both two-stage and end-to-end learning were applied to the “normal-KG” and NR-KG-based methods, respectively.

The results in Table 3 clearly demonstrate the limitations of the “normal-KG” approach, which relies solely on semantics for property prediction and emphasizes the crucial role of numerical data in achieving accurate predictions. The inclusion of numerical data significantly enhances NR-KG-MLP compared to the “normal-KG” approach, underscoring the effectiveness of the cross-modal KG and canonical KG. Furthermore, the incorporation of GCN further improves predictive performance, highlighting the significance of interactions between HEA nodes and their semantic neighbors. Significantly, the end-to-end approach consistently outperforms the two-stage method by leveraging the synergy between KG representation learning and numerical prediction. This approach efficiently learns projection spaces that integrate numerical and semantic aspects, capturing relations within the samples.

4.4.3. Comparative Learning Validation

As the design of the projection prediction loss (PPL) considers the inverse process of the comparative learning loss (CLL), we initially adjust the hyperparameter γ_a in $\mathcal{L}$ and subsequently adjust γ_b to evaluate their impact on model performance. The results in Table 4 show a performance trend where increasing γ_a initially enhances the model’s performance, followed by a decline. Optimal performance for two HEA datasets occurs at $\gamma_a = 0.16$. Similarly, with $\gamma_a = 0.16$, changing γ_b yields similar performance trends, with HEA-HD and HEA-CRD reaching their peaks at $\gamma_b = 0.04$ and $\gamma_b = 0.02$ respectively. These findings highlight that proper PPL and CCL supervision in canonical KG learning can enhance regression tasks.

4.5. Interpretability of NR-KG

4.5.1. Visualization of Semantic Nodes in Canonical KG

We utilized t-SNE dimensionality reduction to visualize the vectors of processing technique nodes in the canonical KG. Nodes were colored based on the classification of processing technique vocabulary by domain experts in the reference materials field, as shown in Fig. 6. Some nodes exhibit spatial patterns consistent with material science expectations. Despite the limited disambiguation and alignment during dataset construction, NR-KG successfully extracts semantic insights from known semantic and numerical information. This potentially explains the enhanced predictive capability of NR-KG when incorporating processing technique data.

4.5.2. Crystal Structure Link Completion

We employed the vectors from the canonical KG to complete the links between HEA nodes and crystal structure nodes, and compared the results with two random methods. The results, presented in Table 5, reveal that even though our model

Table 5: Interpretability of crystal structure nodes vectors and proxy nodes vectors in canonical KG. Results are shown as “mean $\pm$ standard deviation”. Bold cells highlight the best results per metric. “NR-KG($\mathcal{L}_M = 0$)” configures NR-KG training to rely solely on $\mathcal{L}_C$ and $\mathcal{L}_D$, transforming it into a link prediction method.

Datasets Task Methods	HEA-HD			HEA-CRD		
	Link prediction (Testing set)			Link prediction (Testing set)		
	MRR $\uparrow$	MR $\downarrow$	Hit@1 $\uparrow$	MRR $\uparrow$	MR $\downarrow$	Hit@1 $\uparrow$
Uniform sampling random	0.472 $\pm$ 0.031	3.125 $\pm$ 0.233	0.222 $\pm$ 0.039	0.395 $\pm$ 0.011	3.872 $\pm$ 0.105	0.154 $\pm$ 0.017
Distribution sampling random	0.792 $\pm$ 0.016	1.478 $\pm$ 0.036	0.614 $\pm$ 0.031	0.804 $\pm$ 0.013	1.550 $\pm$ 0.044	0.656 $\pm$ 0.025
NR-KG($\mathcal{L}_M = 0$)	0.932 $\pm$ 0.019	1.143 $\pm$ 0.041	0.868 $\pm$ 0.038	0.885 $\pm$ 0.081	1.375 $\pm$ 0.264	0.809 $\pm$ 0.137
NR-KG	0.934$\pm$0.011	1.142$\pm$0.027	0.874$\pm$0.020	0.939$\pm$0.006	1.251$\pm$0.050	0.911$\pm$0.007

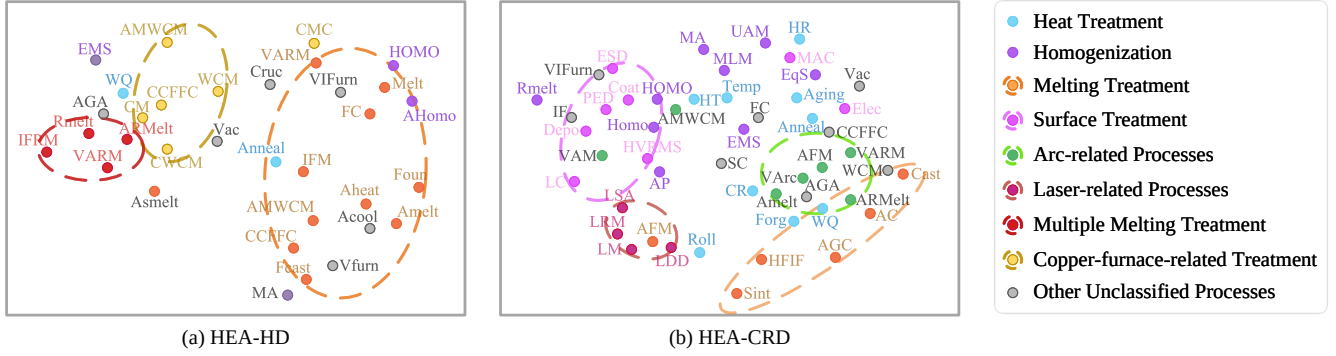


Figure 6: t-SNE visualization of processing technique vectors in the canonical KG, where each point represents a technique. The abbreviation list is in Table A.1.

primarily focuses on material property prediction, it can effectively capture the semantic information of nodes. The accuracy in predicting crystal structures reaches 87.4% and 91.1% in HEA-HD and HEA-CRD. Moreover, in comparison with NR-KG ($\mathcal{L}_M = 0$), we observed that learning material properties enhances the model’s ability to comprehend semantics. This observation underscores NR-KG’s role in bridging the semantic-numerical gap and effectively modeling relationships between samples, thereby extracting more information from small-sample data. The results verify NR-KG’s capability to bridge the semantic-numerical gap and capture inter-sample relationships. This experiment aligns with the empirical understanding in the material field that “process influences structure and structure influences properties” [49], ultimately contributing to the remarkable performance of NR-KG.

4.6. Application Scalability Experiment

To validate the application potential of NR-KG on scientific data, we used publicly available molecular datasets with graph structures on the field of physical chemistry to verify NR-KG.

4.6.1. Baselines

To demonstrate the generalization ability of NR-KG, we comprehensively evaluate its performance on an open dataset in the field of physical chemistry in comparison with molecular-graph-based and molecular-graph pre-training baselines, which encompass the currently recognized state-of-the-art methods, which include GCN [50], GIN [51], MPNN [52], DMPNN [53], CMPNN [54], AttentiveFP [55], N-Gram_{RF} [56], N-Gram_{XGB} [56], PretrainGNN [57], GROVER_{base} [58], GROVER_{large} [58], MolCLR [59], GEM [39], Uni-Mol [40], KANO [11]. For all the baseline methods we report the results taken from the referred papers [39, 40, 11], as shown in Table 6.

Table 6: Molecular experiment results. Results are shown as “mean $\pm$ standard deviation”. Bold cells highlight the best results per metric. Our method, NR-KG, achieved a 54.3% and 22.2% improvement over the second-best algorithm of RMSE. This demonstrates the strong performance of NR-KG on public datasets.

Datasets	FreeSolv	ESOL
No. molecules	642	1,128
Task	Regression (Testing set)	
Methods	RMSE $\downarrow$	
GCN	2.870 $\pm$ 0.135	1.431 $\pm$ 0.050
GIN	2.765 $\pm$ 0.180	1.452 $\pm$ 0.020
MPNN	1.621 $\pm$ 0.952	1.167 $\pm$ 0.430
DMPNN	1.673 $\pm$ 0.082	1.050 $\pm$ 0.008
CMPNN	1.570 $\pm$ 0.442	0.798 $\pm$ 0.112
AttentiveFP	2.073 $\pm$ 0.183	0.877 $\pm$ 0.029
N-Gram _{RF}	2.688 $\pm$ 0.085	1.074 $\pm$ 0.107
N-Gram _{XGB}	5.061 $\pm$ 0.744	1.083 $\pm$ 0.082
PretrainGNN	2.764 $\pm$ 0.002	1.100 $\pm$ 0.006
GROVER _{base}	2.176 $\pm$ 0.052	0.983 $\pm$ 0.090
GROVER _{large}	2.272 $\pm$ 0.051	0.895 $\pm$ 0.017
MolCLR	2.594 $\pm$ 0.249	1.271 $\pm$ 0.040
GEM	1.877 $\pm$ 0.094	0.798 $\pm$ 0.029
Uni-Mol	1.620 $\pm$ 0.035	0.788 $\pm$ 0.029
KANO	1.142 $\pm$ 0.258	0.670 $\pm$ 0.019
NR-KG(Ours)	0.584$\pm$0.235	0.521$\pm$0.042

4.6.2. Experiment Results

The results in Table 6 show that NR-KG outperforms other methods, achieving state-of-the-art performance on molecular datasets. Specifically, on the ESOL dataset and FreeSolv dataset, NR-KG exhibits improvements in RMSE by 22.2% and 54.3%, respectively, compared to the second-best method. Despite not being specifically designed for molecular tasks, NR-KG achieves superior performance by incorporating a simple

adaptation for molecular graph-type data through the addition of a graph feature encoder. It considers molecular fingerprints as semantic information to enhance substructures contained in the molecules, surpassing existing methods in predictive accuracy. This highlights NR-KG’s ability to effectively utilize data from small samples, capturing relevant information from molecular graphs and fingerprints to make accurate predictions.

In summary, the experimental results indicate that the NR-KG model is not only suitable for alloy material datasets but also for predicting molecular properties, showcasing its potential for various types of data.

5. Limitations and Discussion

5.1. Limitations

Although NR-KG effectively utilizes the numerical-semantic information of small-sample data and inter-sample relationship information, yielding promising results in material property prediction tasks, the method’s application is constrained in large-scale datasets due to limitations in the graph convolutional network (GCN) stage. In each iteration, the feature processing of the complete cross-modal KG needs to be considered, leading to increased memory requirements as the dataset and cross-modal KG size grow. This hinders the applicability of NR-KG in large-scale datasets.

5.2. Impact and Prospects

NR-KG provides an innovative solution for integrating numerical and semantic data, introducing new perspectives and tools to the community of machine learning (ML) and scientific data.

In the realm of ML, particularly in KGs and small-sample learning domains:

- NR-KG mines data features and captures information and inter-sample correlations within limited data samples, providing a novel approach for researchers in the field of small-sample ML.
- NR-KG opens up potential avenues for handling more complex modalities and heterogeneous scientific data, expanding the application scope of knowledge graph technology in vertical domains.
- NR-KG has demonstrated potential applications across various scientific domains. In the future, we will further explore the characteristics of data in different scientific domains, aiming to enhance NR-KG to become a more flexible and efficient ML algorithm, contributing to AI for science and interdisciplinary research.

In the community of scientific data:

- In the field of materials science, we propose and publicly release two new High-Entropy Alloy (HEA) datasets, aiming to foster the scientific open-source community and promote the development of high-quality scientific datasets. This work also encourages the collection of diverse datasets in various scientific fields.

- The NR-KG paradigm provides new tools for research in the materials science field. Its excellent interpretability aids in uncovering the complex relationships behind material properties, facilitating in-depth research into phenomena not yet fully understood in materials.

As the first end-to-end KG numerical reasoning method, NR-KG is expected to inspire other tasks in ML, such as cross-modal or small-sample learning. We anticipate that the principles of this algorithm can enhance task performance and provide interpretability to the results. For instance, in common-sense knowledge graph link prediction tasks, more rational embeddings can be obtained through the interaction between numerical and semantic information. In the scientific domain, we hope that NR-KG becomes a valuable tool for scientists, enhancing the accuracy of ML predictions under limited samples and expediting new developments in areas such as the discovery of novel materials.

6. Conclusion

Material property prediction plays a crucial role in various key materials science research areas, including studies on material mechanisms and the discovery of new materials. In this research, we proposed a novel numerical reasoning method of cross-modal KG, NR-KG, for predicting material properties. NR-KG constructs a cross-modal KG to represent material data, effectively capturing both numerical and semantic information. The NR-KG enables end-to-end numerical reasoning within the cross-modal KG. Our proposed NR-KG method outperforms other approaches in property prediction experiments, offering a novel avenue for modeling small-sample scientific data. NR-KG demonstrates its ability to learn material informatics principles effectively even with limited samples, expanding its potential for innovative applications in the scientific domain.

CRedit authorship contribution statement

Guangxuan Song: Writing - Original Draft, Software, Data Curation, Methodology. **Dongmei Fu:** Conceptualization, Supervision. **Zhongwei Qiu:** Writing - Review & Editing. **Zijiang Yang:** Visualization. **Jiaxin Dai:** Validation. **Lingwei Ma:** Data Curation. **Dawei Zhang:** Data Curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code and Datasets Available

The code and HEA datasets will be open-sourced on GitHub.

Acknowledgments

This work was supported by the National Environmental Corrosion Platform of China. We extend our special thanks to the doctoral researchers from Prof. Dawei Zhang’s team at the National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, for their assistance in data collection.

Appendix A. Supplementary Material

Table A.1 serves as the abbreviation list for Fig. 6 in the main text.

Appendix A.1. Experiment Details

Appendix A.1.1. Implemental Details

In 4.4, we examine the influence of two critical hyperparameters, γ_a and γ_b , on NR-KG. The hyperparameter γ in Eq.(6) serves to ensure that negative samples do not constantly yield ineffective loss for $\mathcal{L}_C$. Given the normalization of $\mathbf{e}_i$ in Eq.(4) and $\mathbf{e}_m$ in Eq.(5), $\gamma = 1$ is suggested as a versatile default.

We set the dimension of the canonical KG vectors in NR-KG as $H = 128$, and the graph convolutional network (GCN) has 2 layers to prevent excessive smoothing. We applied dropout = 0.5 in the Property Decoder to mitigate overfitting and utilized the LeakyReLU activation function. The training process consisted of 2000 epochs using the Adam optimizer, and we employed the StepLR learning rate adjustment strategy with gamma = 0.5 and step_size = 1. An early stopping strategy was implemented to enhance generalization, training stopped if the validation loss $\mathcal{L}$ did not decrease for 50 consecutive steps.

For training on the HEA-HD dataset, a learning rate of 0.005 was used with $\gamma = 1$, $\gamma_a = 0.16$, and $\gamma_b = 0.04$ for $\mathcal{L}$. On the HEA-CRD dataset, a learning rate of 0.001 was employed

with $\gamma = 1$, $\gamma_a = 0.16$, and $\gamma_b = 0.02$. The experiments were conducted on an NVIDIA GTX 1080Ti.

For training on the molecule property datasets, we utilize the DeepChem [60] to initialize molecular feature graphs. We adjust learning rates from 0.001, $\gamma_a = 0.2$ and $\gamma_b = 0.04$ for $\mathcal{L}$. All other hyperparameters remained consistent with the experiments outlined on the HEA dataset. The experiments were carried out on an NVIDIA GTX 4090.

Appendix A.1.2. Implemental Details

See Table A.2 for a report on the resource consumption of the main experiments in Sec. 4.3 and 4.6. By effectively utilizing numerical and semantic information, NR-KG achieves state-of-the-art (SOTA) results with a lightweight model of under 0.09M parameters. In terms of time and GPU memory usage, although material property prediction tasks typically have low requirements for real-time performance, NR-KG’s superior data utilization and construction of semantic correlations among data samples within limited data volumes obviate the need for large-scale networks. Consequently, model training can be completed in seconds or minutes, and inference on test set data can be done in milliseconds. With a maximum GPU memory usage of 4.76G, NR-KG can run on most consumer-grade GPUs when dealing with small sample sizes.

Appendix A.2. HEA Datasets Analysis

Appendix A.2.1. Data Collection

The data collection effort was carried out by our team and volunteers from the materials field under the guidance of domain experts. As illustrated in Fig. A.1, we initially compiled existing publicly available materials datasets from [32, 33, 34]. We reviewed the source materials referenced during the data collection of these public datasets and extracted corresponding material processing techniques and crystal structure information. Furthermore, we collected HEA information reported in

Table A.1: Abbreviation of processing techniques. “Abbr.” means “Abbreviation”.

Abbr.	Processing techniques	Abbr.	Processing techniques	Abbr.	Processing techniques
Aging	Aging	Cruc	Crucible	LRM	Laser re-melting
AC	Air casting	Depo	Deposition	LSA	Laser surface alloying
AcCool	Air cooling	EMS	Electromagnetic stirring	MLM	Magnetic levitation melting
AHomo	Alloy homogenization	Elec	Electroplating	MA	Mechanical alloying
Anneal	Annealing	ESD	Electrospark deposition	Melt	Melting
Arc	Arc	EqS	Equilibrium state	MAC	Microplasma arc cladding
AFM	Arc furnace melting	Forg	Forging	PED	Pulsed electric deposition
Afus	Arc fusion	Foun	Foundry	Rmelt	Re-melting
Aheat	Arc heating	FC	Furnace cooling	Roll	Rolling
Amelt	Arc melting	Fcast	Fusion casting	Sint	Sintering
AMWCM	Arc melting in water-cooled copper mold	HT	Heat treatment	SC	Slow cooling
ARMelt	Arc re-melting	HFIF	High-frequency induction heating furnace	Temp	Tempering
Asmelt	Arc smelting	HOMO	Homogeneity	UAM	Ultrasonic atomization method
AGA	Argon gas atmosphere	Homo	Homogenization	Vac	Vacuum
AGC	Argon gas casting	HR	Hot rolling	VArc	Vacuum arc
AP	Atomization process	HVRMS	HV radio frequency magnetron sputtering	VAM	Vacuum arc melting
Cast	Casting	IF	Induction furnace	VARM	Vacuum arc re-melting
CWCM	Casting with copper molds	IFM	Induction furnace melting	Vfurn	Vacuum furnace
Coat	Coating	IFRM	Induction furnace re-melting	VIFurn	Vacuum induction furnace
CR	Cold rolling	LC	Laser cladding	WQ	Water quenching
CCFFC	Copper crucible furnace with furnace cooling	LDD	Laser direct deposition	WCCF	Water-cooled copper crucible furnace
CM	Copper mold	LM	Laser melting	WCM	Water-cooled copper mold
CMC	Copper mold casting				

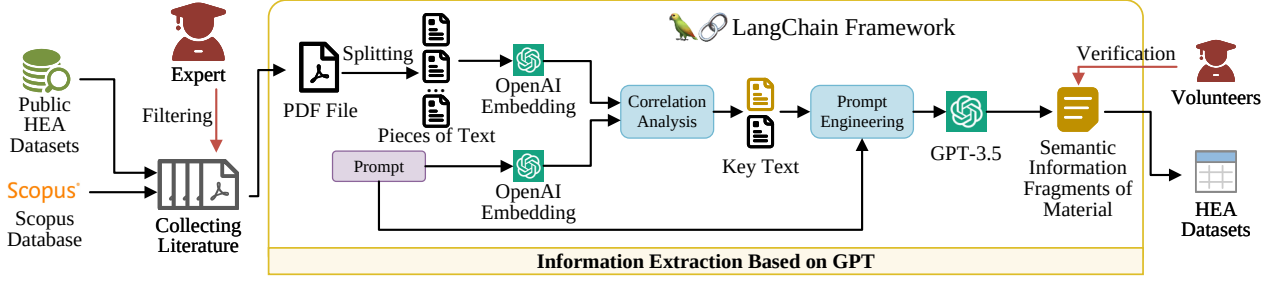


Figure A.1: The semi-automated text extraction framework. We collected data from public HEA dataset sources and high-level papers in the field of HEA from Scopus. Using LangChain, we developed a semi-automated information extraction framework, which was then validated and documented by volunteers from the materials domain.

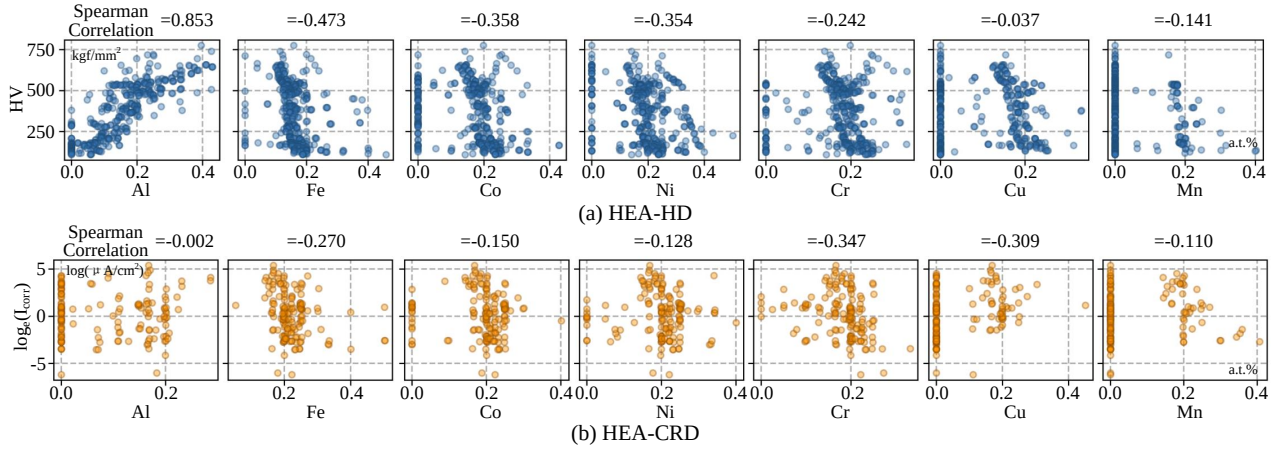


Figure A.2: The correlation between individual element compositions and material properties.

Table A.2: Resource utilization in experiments. Results are reported as “mean $\pm$ standard deviation”. For the HEA-HD and HEA-CRD datasets, 6-fold cross-validation was used, whereas for FreeSolv and ESOL, results are based on three independent data splits, consistent with Sec. 4.3 and 4.6, respectively.

Datasets	Metric		NR-KG
HEA-HD	Training	Parameters	0.074 M
		Total time (Training set size = 214)	13.92 $\pm$ 3.33 s
	Inference	Max GPU memory usage	1.11 $\pm$ 0.00 G
		Total time (Testing set size = 53)	0.04 $\pm$ 0.00 s
HEA-CRD	Training	Parameters	0.077 M
		Total time (Training set size = 92)	8.09 $\pm$ 1.59 s
	Inference	Max GPU memory usage	0.27 $\pm$ 0.00 G
		Total time (Testing set size = 25)	0.01 $\pm$ 0.00 s
FreeSolv	Training	Parameters	0.088 M
		Total time (Training set size = 513)	72.59 $\pm$ 10.79 s
	Inference	Max GPU memory usage	2.99 $\pm$ 0.24 G
		Total time (Testing set size = 65)	0.11 $\pm$ 0.00 s
ESOL	Training	Parameters	0.088 M
		Total time (Training set size = 902)	133.86 $\pm$ 9.82 s
	Inference	Max GPU memory usage	4.76 $\pm$ 0.24 G
		Total time (Testing set size = 114)	0.23 $\pm$ 0.02 s

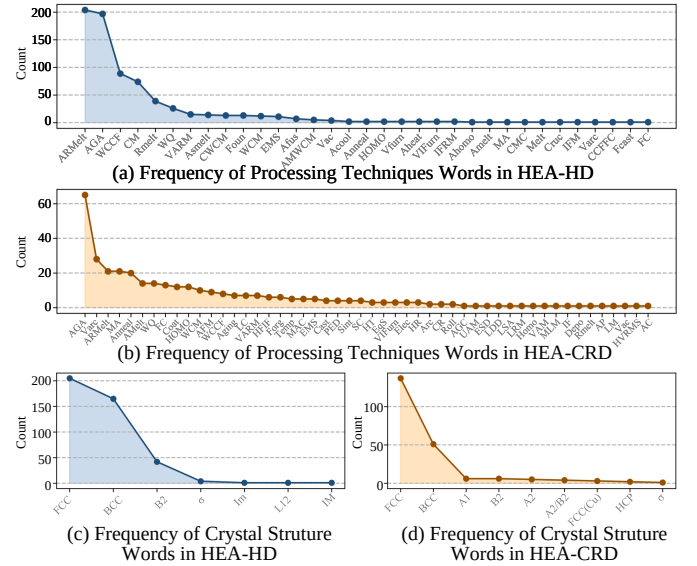


Figure A.3: Frequency of words in HEA-HD and HEA-CRD.

recent high-quality literature in the HEA field. For the aforementioned process, we developed a semi-automated text extraction framework by LangChain [61]: we first parsed PDFs and split paragraphs, then measured the correlation between the PDF paragraphs and prompts using OpenAI Embedding [62] API. Subsequently, paragraphs with sufficient similarity

were fed into GPT-3.5 [35] API after prompt engineering to extract precise sections containing descriptions of HEA processing techniques and crystal structures. Lastly, volunteers from the materials field performed verification and recording. The utilization of GPT-3.5 reduced the workload of literature reading and provided insights for building an automated literature

data collection pipeline in the future.

Appendix A.2.2. Sample Distribution Analysis

The data collection scope encompasses HEAs with Al-Fe-Ni-Co-Cr-Cu-Mn elements. The HEAs we have collected consist of 3-6 elements from the set of Al-Fe-Ni-Co-Cr-Cu-Mn, totaling 7 elements. Consequently, the numerical data contains a substantial number of zero values. This prevalence of zero values is a distinctive characteristic of our HEA dataset.

As shown in Fig. A.2, we depict the correlation between individual element compositions and material properties, along with calculated Spearman correlation coefficients. Analysis reveals that the Al element in the HEA-HD dataset exhibits a noticeable correlation with hardness, while Fe, Co, and Ni also demonstrate certain correlations. Conversely, the correlations

within HEA-CRD are notably weak. These findings align with the comparative experimental results in the full papers, indicating that methods with stronger numerical modeling capabilities yield better coefficient of determination (R^2) results on HEA-HD than on HEA-CRD. The insufficient correlation between elements and performance in HEA-CRD underscores the presence of complex non-linear relationships between material composition and performance, posing challenges for modeling.

Fig. A.3 demonstrates a pronounced long-tail distribution in semantic information, which further contributes to modeling challenges arising from data imbalance. In summary, HEA-HD and HEA-CRD are challenging small-sample material property datasets that incorporate both numerical and semantic information.

References

- [1] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, 6(21):1900808, 2019.
- [2] Jinbo Ma, Jiaxin Dai, Xin Guo, Dongmei Fu, Lingwei Ma, Patrick Keil, Arjan Mol, and Dawei Zhang. Data-driven corrosion inhibition efficiency prediction model incorporating 2d-3d molecular graphs and inhibitor concentration. *Corrosion Science*, page 111420, 2023.
- [3] Jiaxin Dai, Dongmei Fu, Guangxuan Song, Lingwei Ma, Xin Guo, Arjan Mol, Ivan Cole, and Dawei Zhang. Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model. *Corrosion Science*, 209:110780, 2022.
- [4] Cheuk Hei Chan, Mingzi Sun, and Bolong Huang. Application of machine learning for advanced material prediction and design. *EcoMat*, 4(4):e12194, 2022.
- [5] Sherif Abdulkader Tawfik, Olexandr Isayev, Michelle JS Spencer, and David A Winkler. Predicting thermal properties of crystals using machine learning. *Advanced Theory and Simulations*, 3(2):1900208, 2020.
- [6] Chaoshao Gao, Xin Min, Minghao Fang, Tianyi Tao, Xiaohong Zheng, Yangai Liu, Xiaowen Wu, and Zhaohui Huang. Innovative materials science via machine learning. *Advanced Functional Materials*, 32(1):2108044, 2022.
- [7] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3):338–358, 2019.
- [8] Hong Li, Jiajian Lin, Xiaobao Lei, and Tianxia Wei. Compressive strength prediction of basalt fiber reinforced concrete via random forest algorithm. *Materials Today Communications*, 30:103117, 2022.
- [9] Danial Jahed Armaghani, Panagiotis G Asteris, Behnam Askarian, Mahdi Hasanipanah, Reza Tarinejad, and Van Van Huynh. Examining hybrid and single svm models with different kernels to predict rock brittleness. *Sustainability*, 12(6):2229, 2020.
- [10] Shuo Feng, Huiyu Zhou, and Hongbiao Dong. Using deep neural network with small dataset to predict material defects. *Materials & Design*, 162:300–310, 2019.
- [11] Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, pages 1–12, 2023.
- [12] Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3968–3976, 2022.
- [13] Amit Singhal. Introducing the knowledge graph: things, not strings. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>, 2012.
- [14] Agustinus Kristiadi, Mohammad Asif Khan, Denis Lukovnikov, Jens Lehmann, and Asja Fischer. Incorporating literals into knowledge graph embeddings. In *The Semantic Web—ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I*, pages 347–363. Springer, 2019.
- [15] Yanrong Wu and Zhichun Wang. Knowledge graph embedding with numeric attributes of entities. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 132–136, 2018.
- [16] Easo P George, Dierk Raabe, and Robert O Ritchie. High-entropy alloys. *Nature reviews materials*, 4(8):515–534, 2019.
- [17] Xiujian Liu, Pengcheng Xu, Juanjuan Zhao, Wencong Lu, Minjie Li, and Gang Wang. Material machine learning for alloys: Applications, challenges and perspectives. *Journal of Alloys and Compounds*, 921:165984, 2022.
- [18] Luca M Ghiringhelli, Jan Vybiral, Sergey V Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: critical role of the descriptor. *Physical review letters*, 114(10):105503, 2015.
- [19] Jie Zhang, Chen Cai, George Kim, Yusu Wang, and Wei Chen. Composition design of high-entropy alloys with deep sets learning. *npj Computational Materials*, 8(1):89, 2022.
- [20] Dongjiang Niu, Lei Xu, Shourun Pan, Leiming Xia, and Zhen Li. Srrddi: A drug–drug interaction prediction model with substructure refined representation learning based on self-attention mechanism. *Knowledge-Based Systems*, 285:111337, 2024.
- [21] Hui Yu, Jing Wang, Chao Song, and Jian-Yu Shi. Identifying the reaction centers of molecule based on dual-view representation. *Knowledge-Based Systems*, page 111606, 2024.
- [22] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021.
- [23] Valeria Fionda and Giuseppe Pirrò. Learning triple embeddings from knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3874–3881, 2020.
- [24] Guangxuan Song, Dongmei Fu, and Dawei Zhang. From knowledge graph development to serving industrial knowledge automation: A review. In *2022 41st Chinese Control Conference (CCC)*, pages 4219–4226. IEEE, 2022.
- [25] Ren Li, Yanan Cao, Qiannan Zhu, Guanqun Bi, Fang Fang, Yi Liu, and Qian Li. How does knowledge graph embedding extrapolate to unseen data: a semantic evidence view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5781–5791, 2022.
- [26] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
- [27] Ming Zhu, Busra Celikkaya, Parminder Bhatia, and Chandan K Reddy. Latte: Latent type modeling for biomedical entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9757–9764, 2020.
- [28] Alexis Cvetkov-Iliev, Alexandre Allauzen, and Gaël Varoquaux. Relational data embeddings for feature enrichment with background information. *Machine Learning*, pages 1–34, 2023.
- [29] Eda Bayram, Alberto García-Durán, and Robert West. Node attribute completion in knowledge graphs with multi-relational propagation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3590–3594. IEEE, 2021.
- [30] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [31] Tong Li, Xiujian Wang, and Hao Zhong. Cohesive clustering algorithm based on high-dimensional generalized fermat points. *Information Sciences*, 613:904–931, 2022.
- [32] Chen Yang, Chang Ren, Yuefei Jia, Gang Wang, Minjie Li, and Wencong Lu. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Materialia*, 222:117431, 2022.
- [33] Clara Nyby, Xiaolei Guo, James E Saal, Szu-Chia Chien, Angela Y Gerard, Huibin Ke, Tianshu Li, Pin Lu, Christian Oberdorfer, Sarita Sahu, et al. Electrochemical metrics for corrosion resistant alloys. *Scientific data*, 8(1):58, 2021.
- [34] Jie Xiong, San-Qiang Shi, and Tong-Yi Zhang. Machine learning of phases and mechanical properties in complex concentrated alloys. *Journal of Materials Science & Technology*, 87:133–142, 2021.
- [35] OpenAI. gpt-3.5 model. <https://platform.openai.com/docs/models/gpt-3-5>, 2022.
- [36] David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.
- [37] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.
- [38] Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.
- [39] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.
- [40] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.
- [41] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

- [42] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.
- [43] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.
- [44] Uttam Bhandari, Md Rumman Rafi, Congyan Zhang, and Shizhong Yang. Yield strength prediction of high-entropy alloys using machine learning. *Materials Today Communications*, 26:101871, 2021.
- [45] Cheng Wen, Yan Zhang, Changxin Wang, Dezhen Xue, Yang Bai, Stoyichko Antonov, Lanhong Dai, Turab Lookman, and Yanjing Su. Machine learning assisted design of high entropy alloys with desired property. *Acta Materialia*, 170:109–117, 2019.
- [46] Mahmoud Bakr, Junaidi Syarif, and Ibrahim Abaker Targio Hashem. Prediction of phase and hardness of heas based on constituent elements using machine learning models. *Materials Today Communications*, 31:103407, 2022.
- [47] Yao-Jen Chang, Chia-Yung Jui, Wen-Jay Lee, and An-Chou Yeh. Prediction of the composition and hardness of high-entropy alloys by machine learning. *Jom*, 71:3433–3442, 2019.
- [48] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218, 2018.
- [49] Hong Luo, Seok Su Sohn, Wenjun Lu, Linlin Li, Xiaogang Li, Chandrashekar K Soundararajan, Waldemar Krieger, Zhiming Li, and Dierk Raabe. A strong and ductile medium-entropy alloy resists hydrogen embrittlement and corrosion. *Nature communications*, 11(1):3081, 2020.
- [50] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.
- [51] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.
- [52] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [53] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Are learned molecular representations ready for prime time? *arXiv preprint arXiv:1904.01561*, 2019.
- [54] Ying Song, Shuangjia Zheng, Zhangming Niu, Zhang-Hua Fu, Yutong Lu, and Yuedong Yang. Communicative representation learning on attributed molecular graphs. In *IJCAI*, volume 2020, pages 2831–2838, 2020.
- [55] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.
- [56] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.
- [57] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.
- [58] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.
- [59] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.
- [60] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O’Reilly Media, 2019. <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>.
- [61] LangChain. langchain. <https://www.langchain.com/>, 2022.
- [62] OpenAI. Openai embeddings. <https://platform.openai.com/docs/models/embeddings>, 2022.