This article can be cited before page numbers have been issued, to do this please use: M. Q. Ha, D. Le, V. Nguyen, H. Kino, S. Curtarolo and H. Dam, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00400D.

View Article Online
DOI: 10.1039/D5DD00400D

Beyond Interpolation: Integration of Data and AI-Extracted Knowledge for High-Entropy Alloy Discovery

Minh-Quyet Ha,[1] Dinh-Khiet Le,[1] Viet-Cuong Nguyen,[2] Hiori Kino,[3] Stefano Curtarolo,[4,5] and Hieu-Chi Dam[1,6,a)]

[1)] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

[2)] *HPC SYSTEMS Inc., 3-9-15 Kaigan, Minato, Tokyo 108-0022, Japan*

[3)] *Research Center for Materials Informatics, Department of Advanced Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*

[4)] *Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA*

[5)] *Center for Extreme Materials, Duke University, Durham, NC 27708, USA*

[6)] *International Center for Synchrotron Radiation Innovation Smart (SRIS), Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan*

(Dated: 14 December 2025)

Discovering novel high-entropy alloys (HEAs) with desirable properties is challenged by the vast compositional space and the complexity of phase formation mechanisms. Several inductive screening methods that excel at interpolation have been developed; however, they struggle with extrapolating to novel alloy systems. This study introduces a framework that addresses the extrapolation limitation by systematically integrating knowledge extracted from material datasets with expert knowledge derived from scientific literature using large language models (LLMs). Central to our framework is the elemental substitution principle, which identifies chemically similar elements that can be interchanged while preserving desired properties. To model and combine evidence from these multiple sources of knowledge, we employ the Dempster–Shafer theory, which provides a mathematical foundation for reasoning under uncertainty. Our framework consistently outperforms conventional phase selection models that rely on single-source knowledge across all experiments, showing notable advantages in predicting phase stability for compositions containing elements absent from training data. Importantly, the framework effectively complements the strengths of the existing methods. Moreover, it provides interpretable reasoning that elucidates element substitutability patterns critical to alloy stability in HEA formation. These results highlight the framework's potential for knowledge integration, offering an efficient approach to exploring the vast compositional space of HEAs with enhanced generalizability and interpretability.

I. INTRODUCTION

High-entropy alloys (HEAs), also known as multi-principal element alloys (MPEAs), have garnered significant attention owing to their exceptional mechanical properties, thermal stability, and corrosion resistance[1–3]. Typically consisting of five or more principal elements in near-equiatomic ratios, these alloys utilize high-configurational entropy to stabilize single-phase solid solutions[4–6]. However, identifying stable compositions remains a significant challenge due to the vast compositional space and the complex interplay of factors such as mixing entropy, enthalpy, atomic size differences, and electronic structure. These challenges, including exploring expansive design spaces, handling sparse data, and managing uncertainty, represent broader issues in combinatorial materials research, where efficient navigation strategies of compositional possibilities are essential.

A useful framework for understanding this challenge is a decision-making model in which researchers must balance *exploitation* and *exploration*[7,8], as illustrated in Figure 1. Exploitation focuses on well-characterized regions of the design space, having sufficient data for reliable property predictions. This approach supports steady, incremental improvements to existing alloys. In these data-rich regions, uncertainty is primarily *aleatoric*, arising from irreducible variability within the system. Conversely, exploration targets novel regions where data is insufficient for reliable property predictions. These regions introduce higher *epistemic* uncertainty that can be decreased as we collect more data through systematic experimentation. Although exploration bears greater risk, it offers the exciting potential to uncover groundbreaking and fundamentally new alloys with exceptional properties. Achieving an optimal balance between these two strategies is crucial for advancing HEA development.

Data-driven methods have emerged as transformative tools for guiding these exploitation-exploration decisions, enabling the processing of large datasets and streamlining the search for promising HEAs[9–13]. High-throughput approaches, such as CALPHAD[3,14,15], AFLOW[16–18], and Hamiltonian models[19,20], alongside machine learning (ML)[21], have significantly reduced the time and cost associated with evaluating candidate compositions. While

[a)] Electronic mail: dam@jaist.ac.jp

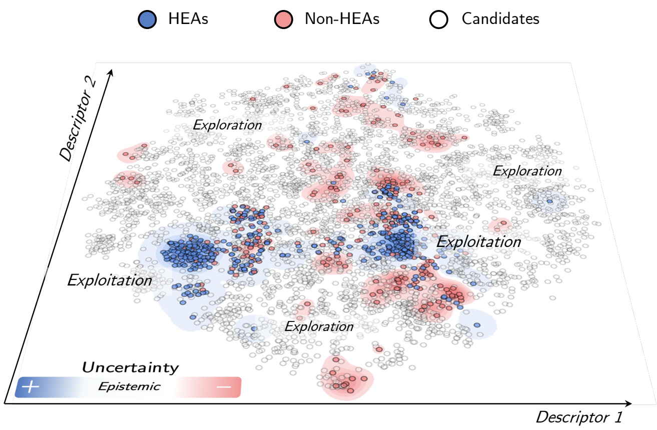


FIG. 1. **Illustration of decision-making scenarios in high-entropy alloy (HEA) discovery.** Colored regions represent well-established areas of the HEA compositional space, characterized by sufficient data suitable for effective exploitation. In contrast, white regions depict unexplored areas with sparse or no existing data, highlighting opportunities for risky yet potentially transformative exploration that could lead to discovering groundbreaking alloys with fundamentally new and exceptional properties. HEAs and Non-HEAs denote alloys that respectively form or do not form a stable high-entropy phase.

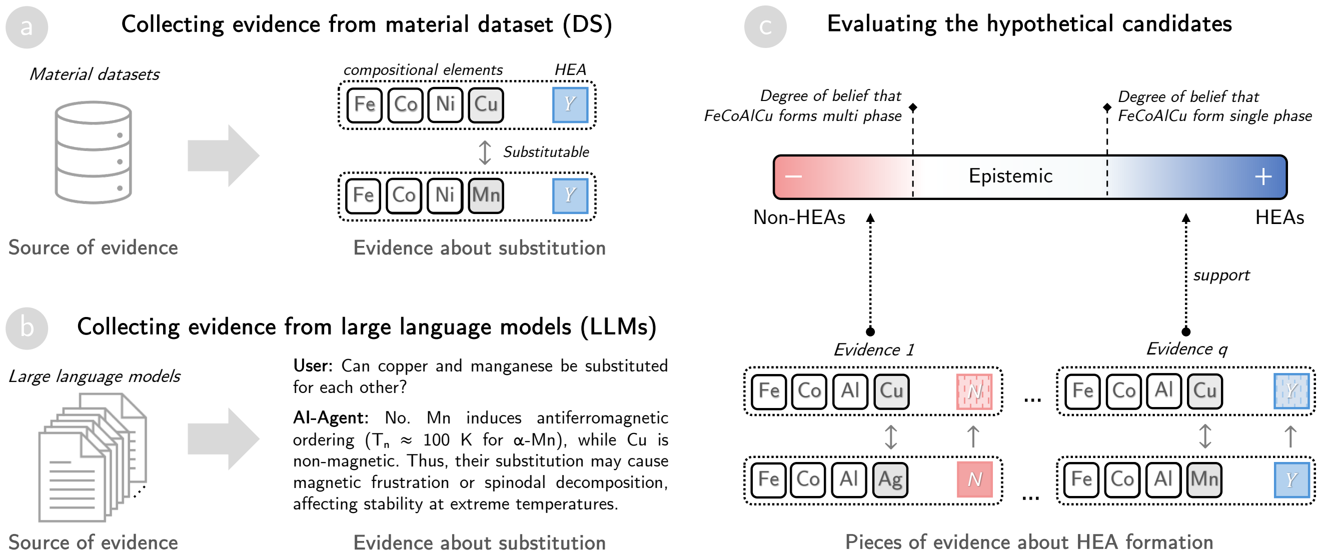
conventional ML models excel at *interpolation*, accurately predicting outcomes for compositions similar to those in the training sets (supporting exploitation), they struggle with *extrapolation* to novel systems, limiting exploration capability[22]. Although careful feature engineering can partially address extrapolation challenges[23], designing features that generalize across vast compositional spaces remains practically difficult[22,24]. This *interpolation–extrapolation* dichotomy needs to be overcome as HEA discovery obviously requires venturing into uncharted territory.

A critical aspect of managing exploration–exploitation balance is uncertainty quantification, which falls into two categories. Epistemic uncertainty arises from incomplete or sparse data and is reducible through targeted information gathering, while aleatoric uncertainty corresponds to intrinsic variability within the system and is irreducible regardless of data volume[25]. Traditional methods, such as Bayesian neural networks, Gaussian processes, and Monte Carlo dropout, are commonly employed to quantify these uncertainties[26,27]. However, they often falter in early-stage materials discovery, where data is sparse or conflicting[28–30].

An alternative framework, the Dempster–Shafer theory[31–33], also known as evidence theory, offers a more flexible means of representing uncertainty. Unlike Bayesian methods, which assign probabilities to individual elements within a set of possibilities (denoted as Ω), evidence theory assigns non-negative weights (summing to one) to subsets of Ω. This enables the explicit representation of ignorance rather than requiring an assumption about a prior probability distribution[25], allowing for nuanced characterization of both epistemic and aleatoric uncertainties. Thus, this framework can guide researchers to specific regions of the compositional space for either efficient exploitation or effective exploration[22,34,35].

However, collecting additional data to reduce epistemic uncertainty is often impractical due to high costs and experimental constraints. Expert knowledge offers a valuable alternative for mitigating this uncertainty. Domain specialists bring insights accumulated across multiple studies and contexts, providing heuristics that extend beyond any single dataset[36–38]. Physics-informed neural networks (PINNs) exemplify one approach to incorporating domain knowledge by embedding a priori physical laws, enabling inference of governing equations from limited observations when those laws are explicit and well-defined[39]. Yet their performance degrades when the underlying physics is only partially understood or key constraints remain unknown. More broadly, expert knowledge often resides in unstructured forms, such as laboratory notebooks, informal rules of thumb, or tacit experience, making its integration with structured, data-driven models a significant challenge.

To bridge this gap, this study introduces a framework that integrates knowledge from material datasets with expert domain knowledge accessed through AI systems–in this implementation, large language models (LLMs) extracting insights from scientific literature–while accounting for inherent uncertainties in each source. This uncertainty-aware integration enables systematic predictions beyond the interpolative boundaries of conventional data-driven methods. Central to our methodology is the *elemental substitution* principle[40,41], a well-established concept in alloy design wherein chemically similar elements can be interchanged while preserving target properties. We treat observed alloy pairs as evidence for substitutability patterns, then consolidate this empirical data with AI-derived insights obtained through state-of-the-art LLMs, including GPT-4o, GPT-4.5, Claude Opus 4, and Grok3. These LLMs leverage documented knowledge from related scientific domains through *knowledge integration* to assess elemental substitutability beyond the training dataset, not by generating information beyond their training corpus. Through Dempster–Shafer theory, the framework systematically models and combines these diverse evidence sources while quantifying both epistemic and aleatoric uncertainties. By providing accurate predictions in well-characterized regions alongside uncertainty-aware guidance for data-sparse spaces, this framework demonstrates–using HEAs as a proof of concept–the viability of materials discovery through uncertainty-aware AI integration.


a
Collecting evidence from material dataset (DS)
Material datasets
Source of evidence
compositional elements
HEA
Fe Co Ni Cu Y
Substitutable
Fe Co Ni Mn Y
Evidence about substitution
b
Collecting evidence from large language models (LLMs)
Large language models
Source of evidence
Evidence about substitution
c
Evaluating the hypothetical candidates
Degree of belief that FeCoAlCu forms multi phase
Degree of belief that FeCoAlCu form single phase
−
Epistemic
+
Non-HEAs
HEAs
support
Evidence 1
Evidence q
Fe Co Al Cu N
Fe Co Al Ag N
Fe Co Al Cu Y
Fe Co Al Mn Y
...
Pieces of evidence about HEA formation

User: Can copper and manganese be substituted for each other?

AI-Agent: No. Mn induces antiferromagnetic ordering ($T_n \approx 100$ K for α-Mn), while Cu is non-magnetic. Thus, their substitution may cause magnetic frustration or spinodal decomposition, affecting stability at extreme temperatures.

FIG. 2. **Hybrid framework integrating Data and AI-extracted Knowledge for high-entropy alloy (HEA) discovery.** (a–b) Schematic depicting the collection of substitutability evidence from a single material dataset (DS) and large language models (LLMs). (c) Schematic illustrating the assessment of hypothetical candidate properties using aggregated evidence derived from substitution-based methods.

II. METHODOLOGY

Each alloy A in the dataset $\mathcal{D}$ is represented by its constituent elements. The property of interest y_A, for any alloy A, can be either HEA or $\overline{HEA}$. Here, HEA denotes alloys that form a stable high-entropy phase (single-phase solid solution), while $\overline{HEA}$ (or Non-HEA) denotes alloys that do not form a stable high-entropy phase (multiphase structures). To determine elemental substitutability, we assess the similarity between different element combinations by adapting *evidence theory*, which models and aggregates diverse pieces of evidence obtained from $\mathcal{D}$. Similarities between objects can manifest in various forms[42]; e.g., pairwise ratings, object sorting, communal associations, substitutability, and correlation. In this study, we specifically focus on the *solid-solution formability* of element combinations and quantify their similarities based on elemental substitutability.

Our approach is intuitively illustrated using the example of element substitutability between Mn and Cu in Figure 2. Suppose we observe from materials datasets that two alloys, FeCoNiCu and FeCoNiMn, both form HEAs. This provides evidence that Cu can substitute for Mn in this context. Meanwhile, consulting domain knowledge through LLMs might reveal that metallurgists consider Cu-Mn pairs as non-substitutable, contributing additional conflicting evidence. Our proposed framework models and combines these independent pieces of evidence using evidence theory, potentially resulting in stronger belief in their substitutability than either source alone would provide. When predicting whether a new alloy, such as FeCoAlCu, forms an HEA, the framework can leverage existing data about FeCoAlMn and the established Cu-Mn substitutability to make informed predictions.

A. Transforming Materials Data to Substitutability Evidence

Consider two alloys, A_i and A_j in $\mathcal{D}$, that share at least one common element. This non-disjoint pair of alloys provides evidence regarding the substitutability between the element combinations:

$$C_t = A_i \setminus (A_i \cap A_j) \quad \text{and} \quad C_v = A_j \setminus (A_i \cap A_j).$$

The intersection $A_i \cap A_j$ serves as the *context* for measuring similarity. If y_{A_i} and y_{A_j} agree (i.e., both are classified as HEA or both as $\overline{HEA}$), we infer that C_t and C_v are substitutable; otherwise, they are non-substitutable, as shown in Figure 2a.

The symmetric substitutability assumption ($C_t \to C_v$ and $C_v \to C_t$ are the same) used in this work represents a context-averaged approximation. While empirically validated for near-equiatomic HEAs, this assumption may limit accuracy for systems with strong directional substitution preferences. However, this symmetric treatment is justified in this study by two factors: first, the limited training data in our data-sparse scenarios makes learning separate directional patterns statistically infeasible; second, for near-equiatomic multi-principal element HEAs characterized by disordered random solid solutions, elements occupy statistically similar local environments, rendering symmetric substitution a physically reasonable first-order approximation.

Evidence for similarity is captured by defining a *frame of discernment*[32] Ω_{sim} = {similar, dissimilar}, encompassing all possible outcomes. The evidence from A_i and A_j is then represented by a *mass function* (or *basic probability assignment*) $m^{C_t,C_v}_{A_i,A_j}$. This mass function assigns non-zero probability to the non-empty subsets of Ω_{sim}, as:

$$m^{C_t,C_v}_{A_i,A_j}(\{\text{similar}\}) = \begin{cases} \alpha, & \text{if } y_{A_i} = y_{A_j}, \\ 0, & \text{otherwise}, \end{cases} \quad (1)$$

$$m^{C_t,C_v}_{A_i,A_j}(\{\text{dissimilar}\}) = \begin{cases} \alpha, & \text{if } y_{A_i} \neq y_{A_j}, \\ 0, & \text{otherwise}, \end{cases} \quad (2)$$

$$m^{C_t,C_v}_{A_i,A_j}(\Omega_{sim}) = 1 - \alpha. \quad (3)$$

Here, the parameter $0 < \alpha < 1$ is determined through an exhaustive search for optimal cross-validation performance, as shown in Supplementary Section 1. Intuitively, $m^{C_t,C_v}_{A_i,A_j}(\{\text{similar}\})$ and $m^{C_t,C_v}_{A_i,A_j}(\{\text{dissimilar}\})$ represent the extent to which alloys A_i and A_j support substitutability or non-substitutability of C_t and C_v. Further, $m^{C_t,C_v}_{A_i,A_j}(\Omega_{sim})$ encodes epistemic uncertainty (i.e., lack of definitive information). The probabilities assigned to these three subsets of Ω_{sim} must sum to 1.

Assuming that we collect q pieces of evidence from $\mathcal{D}$ to compare C_t and C_v, each piece of evidence corresponds to a pair of alloys that generates a mass function m^{C_t,C_v}_i. These q mass functions are combined via *Dempster's rule of combination*[31] to obtain a joint mass function $m^{C_t,C_v}_{\mathcal{D}}$:

$$m^{C_t,C_v}_{\mathcal{D}}(\omega) = \left(m^{C_t,C_v}_1 \oplus m^{C_t,C_v}_2 \oplus \cdots \oplus m^{C_t,C_v}_q\right)(\omega), \quad (4)$$

where $\omega \subseteq \Omega_{sim}$, $\omega \neq \emptyset$ and $\oplus$ denotes the Dempster's rule of combinations, as described in Supplementary Section 2. When no relevant evidence is available, $m^{C_t,C_v}_{\mathcal{D}}$ is initialized with a mass of 1 on $\{\text{similar}, \text{dissimilar}\}$, indicating total uncertainty.

B. Transforming Domain Knowledge to Substitutability Evidence

In addition to evidence collected from material datasets (DS), we focus on evidence derived from domain knowledge, utilizing LLMs to extract insights from a vast corpus of scientific literature. Specifically, we use a set of state-of-the-art LLMs including GPT-4o, GPT-4.5, Claude Opus 4, and Grok3 to assess element substitutability based on expert perspectives within a given domain, as illustrated in Figure 2b. The proposed model evaluates the substitutability of element pairs from the perspective of a domain expert, ensuring that the analysis aligns with established scientific reasoning. To enhance result reliability, we implement a two-step prompting procedure:

- **Question 1:** Do you possess sufficient knowledge or data to evaluate the substitutability of elements C_t and C_v within the context of *[domain knowledge]*?
- **Question 2:** If the answer to the first question is `Yes`, the LLM further rates element substitutability as `High`, `Medium`, or `Low`, based on insights distilled from relevant scientific literature in the given domain.

Detailed prompts used for each LLM are provided in Supplementary File 1. This approach is based on the assumption that, when given clear and structured prompts, these LLMs can simulate expert reasoning across multiple scientific domains. This capability stems from their extensive training on scientific literature, which enables them to provide contextually relevant, domain-specific feedback tailored to the challenges of HEA discovery.

Elemental substitutability is not universal and is property-specific, strongly associated with functionality and applications. For example, substitution for structural stability differs from substitution targeting the magnetic, optical, or mechanical properties. Recognizing this property-specific nature, our framework requires careful domain selection tailored to the target property to ensure accurate predictions. To facilitate the extraction of domain knowledge, we focus on five key scientific domains, including *corrosion science*, *materials mechanics*, *metallurgy*, *solid-state physics*, and *materials science*. These domains are selected due to their critical roles in understanding and optimizing HEAs, specifically tailored for phase stability prediction[5]. Each domain contributes essential insights into different aspects of alloy design.

- **Corrosion science**: This domain examines chemical degradation mechanisms and protective strategies, essential for ensuring long-term durability.
- **Materials mechanics**: This domain investigates mechanical properties such as strength, ductility, and toughness, crucial for structural performance.
- **Metallurgy**: This domain analyzes phase formation, phase diagrams, and microstructure control, offering insights into alloy stability and processing methods.
- **Solid-state physics**: This domain explores atomic-scale interactions, electronic structure, and thermal behavior, all of which influence phase stability and material performance.
- **Materials science**: This domain serves as an integrative field that synthesizes perspectives from the other domains, emphasizing the relationships between composition, structure, properties, and performance to optimize alloy design strategies.

The evidence collected from the LLM for each domain is categorized into one of four outcomes: `High`,

View Article Online
DOI: 10.1039/D5DD00400D

TABLE I. Possible outcomes generated by an LLM for each domain-specific criterion, along with the corresponding mass functions $m^{C_t,C_v}_{\text{LLMs}}(\{\text{similar}\})$, $m^{C_t,C_v}_{\text{LLMs}}(\{\text{dissimilar}\})$, and $m^{C_t,C_v}_{\text{LLMs}}(\{\text{similar}, \text{dissimilar}\})$. Here, $0 < \beta < 1$ indicates our confidence in LLM's response, with determination details provided in Supplementary Section 1.

Q1	**Q2**	$m^{C_t,C_v}_{\text{LLMs}}(\{\text{similar}\})$	$m^{C_t,C_v}_{\text{LLMs}}(\{\text{dissimilar}\})$	$m^{C_t,C_v}_{\text{LLMs}}(\Omega_{sim})$	**Interpretation**
`No`	–	0	0	1	LLM does not provide sufficient domain knowledge
`Yes`	`High`	β	0	$1-\beta$	C_t and C_v are considered *highly* substitutable
`Yes`	`Medium`	$\beta/2$	$\beta/2$	$1-\beta$	C_t and C_v are considered *moderately* substitutable
`Yes`	`Low`	0	β	$1-\beta$	C_t and C_v are considered *poorly* substitutable

`Medium`, `Low`, or `No Knowledge`. Further, these outcomes are mapped to a corresponding mass function denoted as $m^{C_t,C_v}_{\text{LLMs}}$, as shown in Table I. If the LLM indicates `No Knowledge`, then the entire mass is assigned to the set $\{\text{similar}, \text{dissimilar}\}$, reflecting complete epistemic uncertainty. Conversely, if the LLM provides a specific substitutability rating (`High`, `Medium`, and `Low`), then a portion of the mass is allocated to either $\{\text{similar}\}$ or $\{\text{dissimilar}\}$, while the remaining mass is assigned to Ω_{sim} to account for residual uncertainty in the prediction.

Notably, all LLMs (GPT-4o, GPT-4.5, Claude Opus 4, and Grok3) are used as pre-trained models *out-of-the-box* without any fine-tuning, retraining, or in-context literature provision. These models are queried directly through their respective API interfaces using the two-step prompting procedure described above and detailed in Supplementary File 1. The LLMs leverage knowledge from scientific literature encountered during their original pre-training by the respective model developers; we do not modify these models in any way. Each LLM provides independent assessments that are later combined using Dempster-Shafer theory (Section II.C).

C. Combining Evidence from Multiple Sources

In this study, a *source* S refers to an independent knowledge provider that generates evidence about elemental substitutability. Our multi-source framework integrates two kinds of independent sources:

- **DS-source:** A material dataset $\mathcal{D}$ provides empirical evidence by analyzing alloy pairs that differ by element substitution (Section II A). This dataset contains factual observations about the target domain (e.g., which alloy compositions form HEAs).
- **LLM sources:** We query 4 state-of-the-art LLMs (GPT-4o, GPT-4.5, Claude Opus 4, Grok3) across 5 scientific domains (corrosion science, materials mechanics, metallurgy, solid-state physics, materials science), creating $4 \times 5 = 20$ independent knowledge sources (Section II B). Each combination of an LLM and a domain provides documented scientific knowledge from related or similar domains to the target domain.

To integrate substitutability evidence collected from multiple sources, Dempster's rule of combination with a *reliability-aware discounting* step is used[32,43]. Recognizing that substitutability is property-specific and different sources capture different aspects of elemental substitutability, our framework implements an adaptive mechanism that evaluates each source's relevance to the target property. This reliability-aware discounting automatically assigns higher weights to sources that align well with the specific property being predicted while suppressing sources that capture irrelevant substitutability criteria, thereby preventing inappropriate knowledge integration.

For each source S, we compute a dataset-specific discount factor as:

$$\gamma_S = \text{disc}\left(m^{C_t,C_v}_S, \mathcal{D}\right) \in [0,1], \tag{5}$$

where disc(.) quantifies how well the substitutability evidence collected from source S generalizes to the alloy properties in $\mathcal{D}$. The reliability of each source is assessed using the macro-averaged F1 score with 10-fold cross-validation. For instance, if a source S has historically demonstrated accurate predictions on alloys similar to those in $\mathcal{D}$, we assign γ_S a value closer to 1. Conversely, if S performs poorly or unpredictably for alloys in $\mathcal{D}$, γ_S is reduced accordingly.

The original mass function m^{C_t,C_v}_S for source S is then modified by incorporating the discount factor γ_S, leading to an adjusted function ${}^{\gamma_S}m^{C_t,C_v}_S$:

$$\begin{aligned}
{}^{\gamma_S}m^{C_t,C_v}_S(\{\text{similar}\}) &= \gamma_S \times m^{C_t,C_v}_S\big(\{\text{similar}\}\big),\\
{}^{\gamma_S}m^{C_t,C_v}_S(\{\text{dissimilar}\}) &= \gamma_S \times m^{C_t,C_v}_S\big(\{\text{dissimilar}\}\big),\\
{}^{\gamma_S}m^{C_t,C_v}_S(\Omega_{sim}) &= 1 - \gamma_S + \gamma_S \times m^{C_t,C_v}_S\big(\Omega_{sim}\big).
\end{aligned} \tag{6}$$

This redistribution shifts mass from definitive conclusions $\{\text{similar}\}$ and $\{\text{dissimilar}\}$ to the ambiguous set $\{\text{similar}, \text{dissimilar}\}$, thereby encoding epistemic uncertainty for less reliable sources. Therefore, when all mass functions are subsequently merged using Dempster's rule, less credible sources exert a weaker influence on the final decision.

Assuming p sources $\{S_1, S_2, \ldots, S_p\}$, the substitutability evidence gathered from them is aggregated using Dempster's rule of combination:

$$m^{C_t,C_v}(\omega) = \left({}^{\gamma_{S_1}}m^{C_t,C_v}_{S_1} \oplus {}^{\gamma_{S_2}}m^{C_t,C_v}_{S_2} \oplus \cdots \oplus {}^{\gamma_{S_p}}m^{C_t,C_v}_{S_p}\right)(\omega), \tag{7}$$

View Article Online
DOI: 10.1039/D5DD00400D

where ω denotes non-empty subsets of Ω_{sim}. The rule iteratively integrates evidence while normalizing conflicts (such as empty-set intersections arising from contradictory sources). This approach preserves diverse insights, from data-driven correlations to LLM-derived domain knowledge, while mitigating the influence of unreliable sources. Critically, when evidence about substitutability is insufficient or conflicting, Dempster's rule of combination assigns high mass to $m^{C_t,C_v}(\Omega_{sim})$, explicitly signaling uncertainty rather than forcing confident predictions. This naturally prevents overfitting in data-sparse scenarios common in materials discovery.

Similar analyses are conducted for all pairs of element combinations, resulting in a symmetric matrix M, where $(M[t,v] = M[v,t] = m^{C_t,C_v}(\{\text{similar}\}))$.

D. Evaluating Hypothetical Candidates by Analogy-Based Inference

To predict whether a *new* alloy A_{new} is likely to form an HEA, we employ a substitution-based inference approach utilizing the similarity matrix M. The process begins with a known alloy A_k, labeled y_{A_k}, and identifies the subset $C_t \subset A_k$ that, when replaced by C_v, generates A_{new} (Figure 2 c). If C_t and C_v are deemed substitutable, then $y_{A_{new}}$ is more likely to match y_{A_k}; conversely, if they are dissimilar, $y_{A_{new}}$ may differ.

We formalize this inference using a frame of discernment[32] $\Omega_{HEA} = \{\text{HEA}, \overline{\text{HEA}}\}$ and define a mass function $m^{A_{new}}_{A_k, C_t \leftarrow C_v}$ to model the evidence collected from A_k and the substitution of C_t, for C_v, denoted as $C_t \leftarrow C_v$. This mass function distributes belief among $\{\text{HEA}\}$, $\{\overline{\text{HEA}}\}$, or $\{\text{HEA}, \overline{\text{HEA}}\}$ according to the similarity $M[t,v]$ and the label of A_k as:

$$m^{A_{new}}_{A_k, C_t \leftarrow C_v}(\{\text{HEA}\}) = \begin{cases} M[t,v], & \text{if } y_{A_k} = \text{HEA}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

$$m^{A_{new}}_{A_k, C_t \leftarrow C_v}(\{\overline{\text{HEA}}\}) = \begin{cases} M[t,v], & \text{if } y_{A_k} = \overline{\text{HEA}}, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

$$m^{A_{new}}_{A_k, C_t \leftarrow C_v}(\Omega_{HEA}) = 1 - M[t,v]. \tag{10}$$

Here, the probability mass assigned to $\{HEA\}$ and $\{\overline{HEA}\}$ reflects the confidence levels with which A_k and the substitution of C_v for C_t support the probabilities that A_{new} is or is not an HEA, respectively. The mass assigned to subset $\{HEA, \overline{HEA}\}$ represents epistemic uncertainty, signifying cases where the available evidence does not provide definitive information regarding the properties of A_{new}. The total probability mass assigned to all three non-empty subsets of Ω_{HEA} is constrained to sum to 1, ensuring a consistent probabilistic framework. An illustrative example employing the Dempster-Shafer theory for the evaluation of hypothetical candidates is provided in Supplementary Section 3.

We assume that multiple pieces of evidence can be collected, each derived from a distinct pair of host alloy A_{host} and substitution pair $C_t \leftarrow C_v$, for a new alloy candidate A_{new}. These individual pieces of evidence are systematically combined using Dempster's rule of combination to generate a final mass function $m^{A_{new}}$. This function integrates all available analogies, resolving potential inconsistencies and contradictions among the sources. The resulting combined evidence offers a coherent assessment, aiding in informed decision-making regarding whether further resource-intensive experiments are necessary to validate the HEA formation ability of A_{new}.

III. EXPERIMENTAL SETTING

In this section, we present the design of experiments, which assess both the *predictive capability* and *interpretability* of our proposed method. Additionally, we provide comparisons against alternative approaches, including single-source evidential methods and other data-driven classifiers.

A. Datasets

Experiments are conducted considering four computational datasets of quaternary alloys, one experimental dataset of quaternary alloys, and one experimental dataset of quinary high-entropy borides (HEB), summarized in Table II. HEBs are single-phase ceramics containing multiple transition metal cations randomly distributed on the metal sublattice of a boride structure, offering unique combinations of metallic and ceramic properties[44]. Despite different bonding mechanisms, HEBs exhibit similarly high elemental selectivity as HEAs–boron's restrictive bonding requirements create stringent constraints on metal selection, analogous to the selective substitutability patterns in metallic HEAs, making them suitable for testing our framework's core principle of managing uncertainty in highly selective multicomponent systems.

- $\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350\text{K}}$: These computational datasets include *all possible quaternary* alloys generated from a set of 26 elements: Fe, Co, Ir, Cu, Ni, Pt, Pd, Rh, Au, Ag, Ru, Os, Si, As, Al, Re, Mn, Ta, Ti, W, Mo, Cr, V, Hf, Nb, and Zr. The stability of these alloys is predicted using methods proposed by Chen *et al.*[45] at two different temperatures: $0.9\,T_m$ (approximately 90% of the melting temperature T_m of the alloy) and $1350\,(K)$. These predictions are obtained via a high-throughput computational workflow, which employs a regular-solution model[46,47] using binary interaction param-

View Article Online
DOI: 10.1039/D5DD00400D

TABLE II. Summary of alloy datasets used in evaluation experiments. No. alloys: Total number of alloys present in each dataset. No. positive label: Number of alloys classified as forming HEA phases in datasets $\mathcal{D}_{0.9Tm}$ and $\mathcal{D}_{1350\text{K}}$, the number of alloys exhibiting non-zero magnetization in $\mathcal{D}_{\text{Mag}}$, and the number of alloys with a non-zero Curie temperature in $\mathcal{D}_{T_C}$. The percentage values in parentheses represent the proportion of positive labels within each dataset.

Dataset	**No. alloys**	**Physical properties**	**Positive label**	**No. positive label**
$\mathcal{D}_{0.9T_m}$	14,950 quaternary alloys	Stability	HEA	4,218 (28%)
$\mathcal{D}_{1350\text{K}}$	14,950 quaternary alloys	Stability	HEA	1,402 (9%)
$\mathcal{D}_{\text{Mag}}$	5,968 quaternary alloys	Magnetization (T)	Magnetic	2,428 (41%)
$\mathcal{D}_{T_C}$	5,968 quaternary alloys	Curie temperature (K)	Non-zero Curie Temperature	2,355 (39%)
$\mathcal{D}^{\text{exp}}_{\text{HEA}}$	55 quaternary alloys	Stability	HEA	40 (73%)
$\mathcal{D}^{\text{exp}}_{\text{HEB}}$	19 quinary alloys-borides	Stability	HEB	15 (79%)

eters derived from *ab initio* density functional theory (DFT) to compute and compare Gibbs free energies of solid solutions against competing intermetallic phases[16–18].

- $\mathcal{D}_{Mag}$ and $\mathcal{D}_{T_C}$: These computational datasets comprise 5,968 quaternary high-entropy alloys (HEAs)[35], each formed by selecting four elements from a set of 21 transition metals: Fe, Co, Ir, Cu, Ni, Pt, Pd, Rh, Au, Ag, Ru, Os, Tc, Re, Mn, Ta, W, Mo, Cr, V, and Nb. Their magnetizations ($\mathcal{D}_{Mag}$) and Curie temperatures ($\mathcal{D}_{T_C}$) in the body-centered cubic (BCC) phase are computed using the Korringa–Kohn–Rostoker coherent approximation method[48]. These datasets are derived from an original pool of $147,630$ equiatomic quaternary HEAs.

- $\mathcal{D}^{\text{exp}}_{\text{HEA}}$: The experimental dataset includes 55 experimentally verified quaternary HEAs from peer-reviewed publications[45,49,50]. The dataset includes both HEA (40 alloys) and non-HEA (15 alloys) compositions, providing balanced representation for validation.

- $\mathcal{D}^{\text{exp}}_{\text{HEB}}$: The experimental dataset includes 19 experimentally verified quinary HEBs from peer-reviewed publications[44]. The dataset includes 15 quinary systems forming HEB.

B. Design of experiments

We begin by verifying the reliability of the elemental substitutability knowledge queried from large language models (LLMs). Specifically, we compare the LLM-derived substitutability knowledge with the well-established Hume–Rothery criteria for elemental substitution.

With that reliability confirmed, we turn to predictive capability. Two experiments on four computational datasets serve as the framework's proving ground to evaluate predictive capability of our proposed framework: (1) Cross-validation on quaternary alloys, assessing performance with randomly partitioned training sets (1%-30% of data) to determine how effectively LLM-derived knowledge aligns with material-specific relationships across different data availability scenarios, with particular focus on data-limited conditions; and (2) Extrapolation on quaternary alloys, simulating real discovery scenarios by excluding alloys containing a specific element from training and evaluating performance on compositions that incorporate this previously unseen element. These computational datasets, free from experimental bias and large enough for robust statistics, provide the controlled environment needed for *framework development.*

To benchmark our multi-source method, we compare its predictive performance against two baseline approaches.

- **Single-source methods:** These methods rely exclusively on one source of evidence, either a material dataset or domain knowledge derived from only one LLM from the set of state-of-the-art models under investigation.

- **Traditional classification method:** We employ logistic regression (LR)[51].

Hyper-parameters of these methods are tuned via systematic grid search, as detailed in Supplementary Section 1. Hereinafter, we define models employing the evidential method (based on the Dempster–Shafer theory) as follows: models trained solely on material datasets are termed DS-source models; those leveraging evidence from LLMs are termed LLM-source models; and those integrating both sources are termed multi-source models. Notably, the LLM-source models are obtained by combining 20 independent sources–each of the 4 LLMs (GPT-4o, GPT-4.5, Claude Opus 4, Grok3) queried across 5 scientific domains–through Dempster-Shafer theory (Section II C). The multi-source model further integrates this combined LLM-source with the DS-source using the same framework. Models utilizing logistic regression and support vector machines are referred to as LR-based model.

To assess the real-world applicability of our framework, we next validate its predictive performance on experimentally verified alloys. This validation examines whether the proposed framework can accurately predict phase stability for experimentally synthesized alloys. Our framework integrates LLM-derived knowledge with

View Article Online
DOI: 10.1039/D5DD00400D

TABLE III. Confusion matrix comparing LLM consensus predictions with Hume–Rothery rules for 351 element pairs considered in this study.

		Hume–Rothery rules		
		Substitutable	Non-substitutable	Total
LLMs	Substitutable	33 pairs (*True positive*)	45 pairs (*False positive*)	78 pairs
	Non-substitutable	4 pairs (*False negative*)	269 pairs (*True negative*)	273 pairs
	Total	37 pairs	314 pairs	351 pairs

substitutability patterns extracted from computational datasets. This reflects real-world scenarios where researchers must consider all available knowledge to fill the gaps raised by limited experimental data before selecting candidates for expensive synthesis. Finally, after evaluating the predictive performance across all settings, we analyze the element substitutability patterns captured using the multi-source approach to gain deeper insights into the underlying HEA formation mechanisms of quaternary alloys.

C. Materials descriptors

Descriptors, which are the representation of alloys, play a crucial role in building a recommender system to explore potential new HEAs. In this research, the raw data of alloys is represented in the form of element combinations. Several descriptors have been studied in materials informatics to represent the compounds[52]. To employ the data-driven approaches for this work, we applied compositional descriptor[53] and binary elemental descriptor.

Compositional descriptors represent each alloy through 135 features derived from 15 atomic properties of constituent elements. These properties include structural parameters (*atomic number*, *mass*, *period*, and *group*), electronic characteristics (*first ionization energy*, *second ionization energy*, *Pauling electronegativity* and *Allen electronegativity*), size factors (*van der Waals*, *covalent*, and *atomic radii*), and thermophysical properties (*melting point*, *boiling point*, *density*, *specific heat*). For each atomic property, we calculate statistical numbers, including mean, standard deviation, and pairwise covariances across the alloy's elements, to represent the alloy. The compositional descriptors can be applied not only to crystalline systems but also to molecular systems. However, the descriptors cannot easily distinguish alloys with different numbers of constituent elements, because they treat the atomic properties as statistical distributions. Therefore, the descriptors cannot be applied when extrapolating to alloys with a different number of components.

Binary elemental descriptors use binary encoding to indicate element presence (1) or absence (0) in an alloy. The number of binary elemental descriptors corresponds to the number of element types included in the training data. In this study, the binary elemental descriptors are used to represent the alloys in the DS-source, LLM-source, and multi-source models. In contrast, the compositional descriptors are applied for the LR-based model.

IV. RESULTS AND DISCUSSIONS

A. Reliability Assessment of LLM-Based Elemental Substitutability Knowledge

Verifying the reliability of large language model (LLM) responses is a prerequisite for trusting downstream predictions. We therefore validate element-substitutability knowledge extracted from LLM queries against the empirical Hume–Rothery rules[54], which are a set of basic rules for predicting elemental substitution. These rules stipulate that elements readily substitute in solid solutions when: (i) atomic radius mismatch is lower than 15%, (ii) they share similar crystal structures and valence states, and (iii) they have similar electronegativity. When electronegativity differences exceed critical thresholds, metals typically form intermetallic compounds rather than solid solutions. For this validation, we use an electronegativity difference threshold of 0.55. For valency comparison in metallic alloy systems, we consider the effective valency[55] (number of electrons effectively contributing to metallic cohesion). While most metals exhibit a single characteristic valency, certain transition metals (e.g., Fe, Co, Mn, Cr) can exhibit multiple effective valencies in different alloy environments. In our analysis, two elements are considered to have similar valency if they share at least one common valence state.

We aggregated substitutability assessments from four LLMs, including Grok3, Claude Opus 4, GPT-4o, and GPT-4.5, for 351 element pairs using our DST framework. Each pair is classified as substitutable if the combined belief for substitutability exceeds that for non-substitutability. Comparison against Hume–Rothery predictions reveals strong alignment: 86% of element pairs show identical classifications with high recall rates for substitutable labels and high precision for non-substitutable labels, as shown in Table III. Specifically, 33 of 37 pairs (89%) deemed substitutable by Hume–Rothery rules are correctly identified by LLMs, while

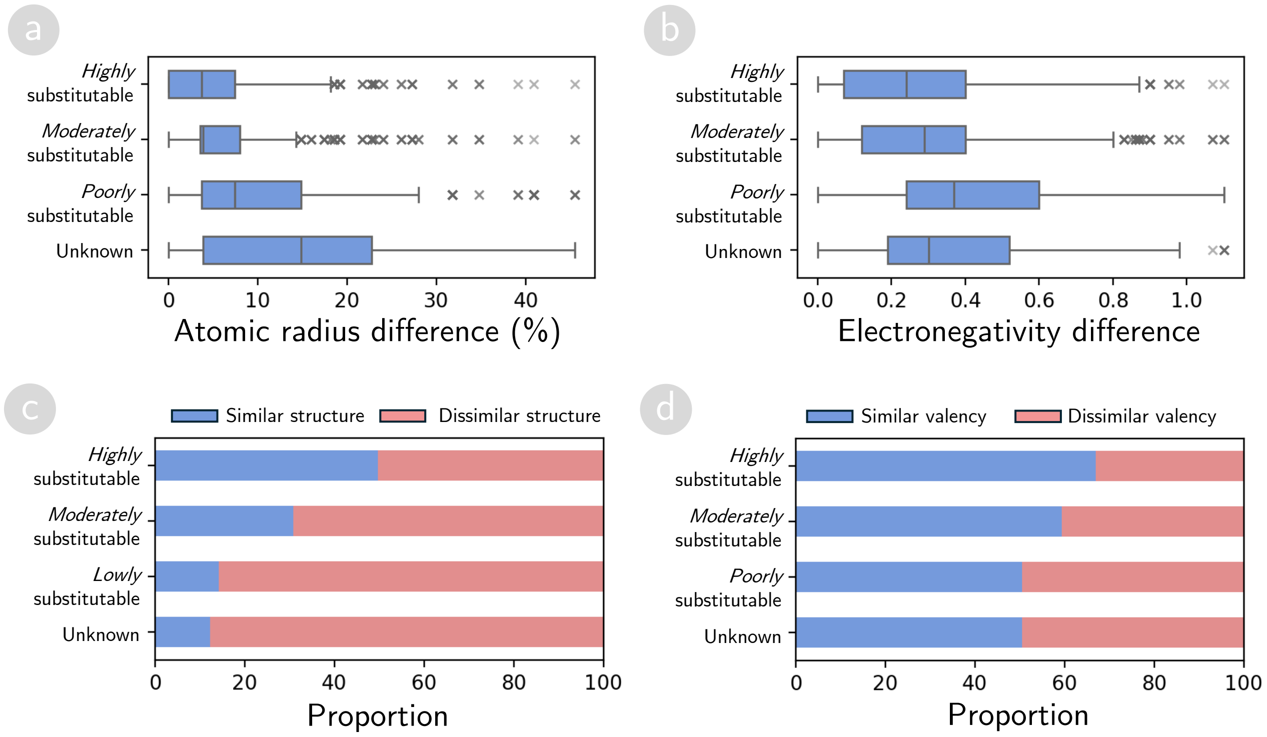


FIG. 3. **Validation of LLM-extracted substitutability against Hume–Rothery rules.** (a, b) Distribution of atomic radius differences (a) and electronegativity differences (b) for element pairs categorized by LLM-predicted substitutability levels (highly, moderately, and poorly substitutable, plus unknown). Box plots show median, interquartile range, and outliers. (c, d) Proportions of element pairs with similar versus dissimilar crystal structures and valency, grouped by substitutability levels.

269 of 273 pairs classified as non-substitutable by LLMs matched Hume–Rothery rules, achieving a precision of 99%.

The 14% misalignment consists entirely of cases where LLMs identify additional substitutable pairs beyond the traditional Hume–Rothery criteria. Among the 45 misaligned pairs, most satisfy the size and electronegativity requirements but exceed traditional thresholds for valency or crystal structure differences. Remarkably, experimental validation supports these context-specific predictions: 14 of these pairs have been confirmed to form single-phase binary systems[56], as shown in Supplementary Table 3. Additionally, Cr and Nb differ in valence electron counts (Cr: 6, Nb: 5), placing them outside general substitutability criteria. However, when incorporated into quaternary systems, they demonstrate successful substitution–Cr in quaternary system Cr-Al-Ti-V can be replaced by Nb (forming Nb-Al-Ti-V), and similarly in Cr-Ta-Ti-V and Nb-Ta-Ti-V systems, both form stable single-phase BCC structures.

This asymmetric difference reflects a fundamental distinction between general rules and context-specific knowledge. The Hume-Rothery rules, developed through careful empirical observation, provide general guidelines with well-defined thresholds (e.g., 15% for radius difference) that have successfully guided alloy design for decades. These universal criteria ensure high reliability across diverse alloy systems. In contrast, LLMs capture context-dependent substitutability documented in materials literature[57], in which specific processing conditions, alloy compositions, or applications enable successful substitution despite exceeding general thresholds. LLMs integrate knowledge from documented experimental systems across material families for general substitutability assessment, explaining why they complement conservative Hume-Rothery rules with context-specific insights. Detailed analysis of all 45 pairs with experimental validation status is provided in Supplementary Table 3.

Figure 3 analyzes in detail the alignment of LLM's response with each criterion of substitutability from Hume–Rothery rules. Element pairs that LLMs identified as highly substitutable exhibit significantly lower atomic radius differences and electronegativity differences compared to pairs identified as poorly substitutable, as shown in Figure 3(a-b). Additionally, highly substitutable pairs predominantly share similar crystal structures and valencies, while poorly substitutable pairs rarely do as shown in Figure 3(c-d).

View Article Online
DOI: 10.1039/D5DD00400D

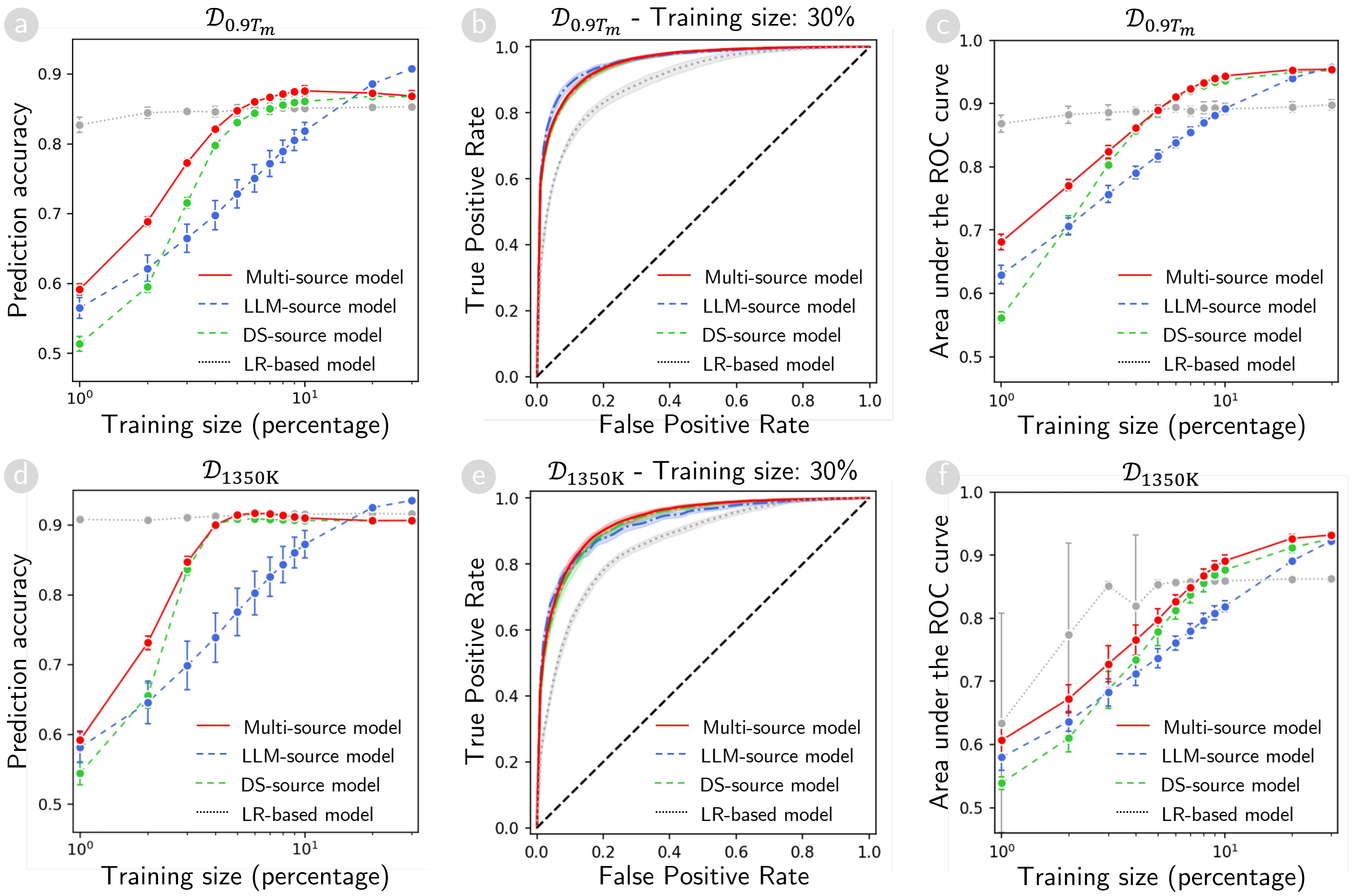


FIG. 4. **Predictive capability evaluation via cross-validation on quaternary-alloy datasets $\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350K}$.** (a, d) Classification accuracy of the multi-source, single-source, and LR-based models on two quaternary alloy datasets $\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350K}$. (b, e) Receiver operating characteristic (ROC) curves for the same models at a 30% training-set size on these datasets. (c, f) Area under the ROC curves (AUC) for each model across different training-set sizes, providing an overall measure of discriminative performance. In all subplots, red lines indicate the multi-source model (using both DS and LLM sources), green and blue lines represent single-source models (using either DS or LLM sources), and gray lines represent the LR-based model.

B. Cross-Validation Analysis of Multi-Source Knowledge Integration

For the experiment, we systematically vary the training set size from 1% to 30% of each quaternary-alloy dataset, incrementing by 1% up to 10%, followed by steps of 20% and 30%. The variation enables the assessment of how different methods handle data scarcity versus moderate availability.

Figures 4(a,d) and 5(a,d) show the classification accuracy of the single-source, multi-source, and LR-based models on the four datasets. At smaller training sizes (approximately 1%–10%), the LR-based model achieves the highest overall accuracy, outperforming evidential models, which explicitly model element substitutability to predict alloy properties. Among the evidential models, single-source LLM models initially outperform DS-source models, attributed to LLM-derived domain-specific insights that assist in mitigating data limitations. However, multi-source models remain competitive and sometimes achieve the highest accuracy among evidential models, even with limited data. As the training size exceeds 10%, DS-source models exhibit superior performance on the magnetization and Curie temperature datasets while achieving comparable accuracy to LLM-source models on alloy stability datasets. Conversely, the accuracy of LR-based models plateaus and is eventually outperformed by evidential models. These findings underscore the importance of incorporating LLM-based, DS-source, or multi-source knowledge to improve quaternary-alloy property predictions.

Although prediction accuracy provides a convenient single-metric overview, it relies on a fixed classification threshold (typically 0.5), which may not be optimal for imbalanced datasets, where HEAs (positive class) are relatively rare. Under these conditions, LR-based models may serve effectively at extremely small training sizes when they effectively predict the dominant (Non-HEA) class by default, thereby inflating accuracy. However,

View Article Online
DOI: 10.1039/D5DD00400D

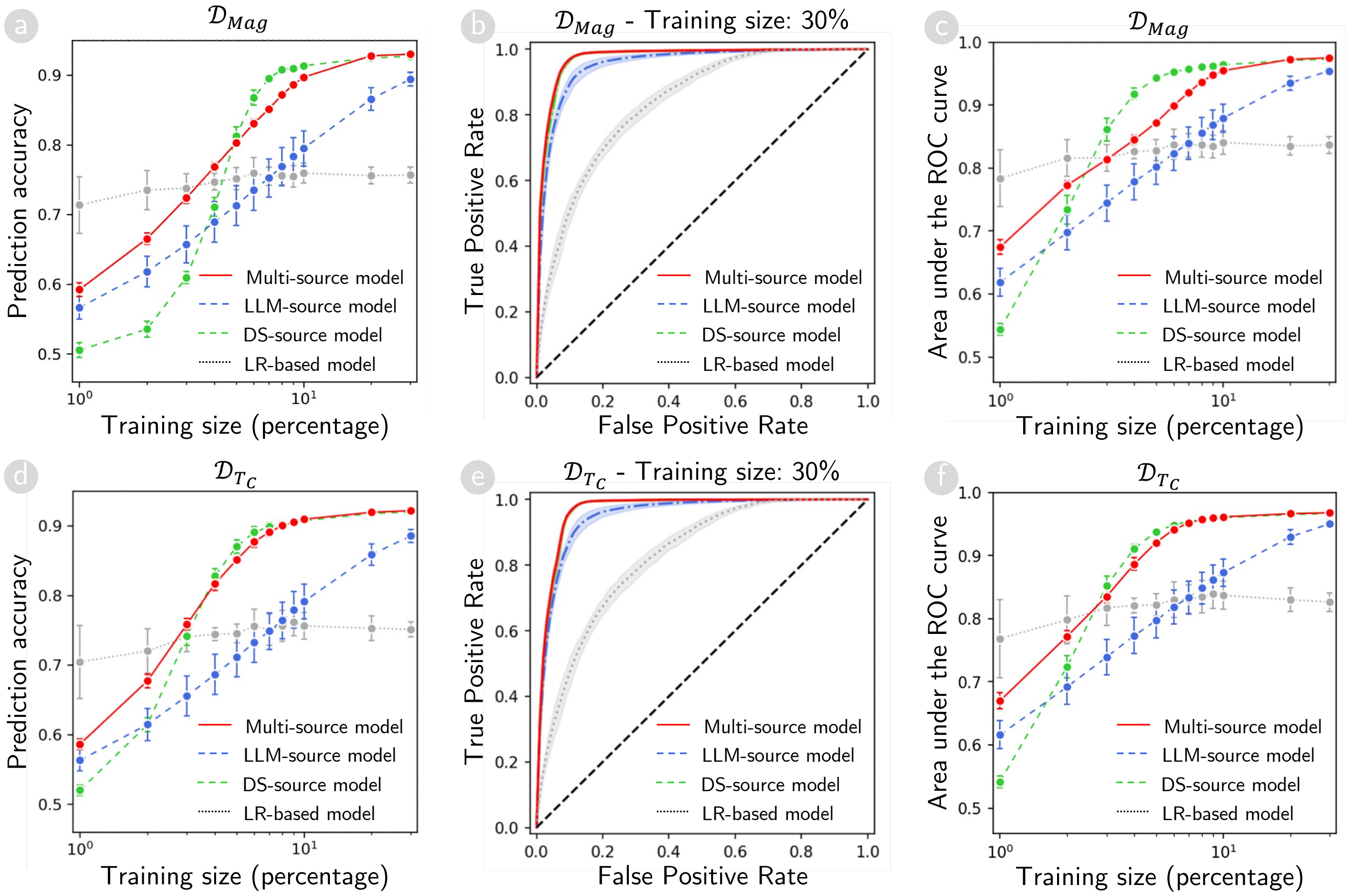


FIG. 5. **Predictive capability evaluation via cross-validation on quaternary-alloy datasets $\mathcal{D}_{\mathrm{Mag}}$ and $\mathcal{D}_{T_C}$.** (a, d) Classification accuracy of the multi-source, single-source, and LR-based models on two quaternary alloy datasets $\mathcal{D}_{\mathrm{Mag}}$ and $\mathcal{D}_{T_C}$. (b, e) Receiver operating characteristic (ROC) curves for the same models at a 30% training-set size on these datasets. (c, f) Area under the ROC curves (AUC) for each model across different training-set sizes, providing an overall measure of discriminative performance. In all subplots, red lines indicate the multi-source model (using both DS and LLM sources), green and blue lines represent single-source models (using either DS or LLM sources), and gray lines represent the LR-based model.

this approach fails to address scenarios where different types of misclassifications (false positives versus false negatives) incur different costs.

To effectively capture these trade-offs under dynamic thresholds, we analyze receiver operating characteristic (ROC) curves across the four datasets, which illustrate variations in true positive rate (TPR) and false positive rate (FPR) of each model across all possible decision boundaries. Figures 4(b,e) and 5(b,e) depict the ROC curves for the multi-source models, LLM-source models, DS-source models, and LR-based models at a 30% training size. Overall, the multi-source and DS-source models exhibit comparable ROC performance and outperform the other models. The LLM-source models achieve results comparable to the best ones on the alloy stability datasets $\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350\mathrm{K}}$ but lag behind DS-source models on the magnetization and Curie temperature datasets $\mathcal{D}_{\mathrm{Mag}}$ and $\mathcal{D}_{T_C}$. Therefore, knowledge collected from the five considered research domains may not fully capture the magnetic and thermal properties reflected in those datasets. Meanwhile, the LR-based models consistently show the lowest performance across all four datasets.

To further assess the ROC performance of each model at different training sizes, we analyze the AUC distribution from 1% to 30% training data, as shown in Figures 4(c,f) and 5(c,f). When the training set is extremely small, LLM-based models generally attain an early advantage, presumably because domain insights compensate for limited alloy observations. However, as data accumulates, DS-source models typically outperform LLM-source models, suggesting that direct data-driven cues from quaternary-alloy datasets become increasingly decisive. In contrast, multi-source models maintain robust performance across all training sizes, benefitting from their ability to merge domain-specific substitutability insights with empirical data. Multi-source models leverage complementary evidence, enabling an effective balance between TPR and FPR. On stability datasets $\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350\mathrm{K}}$, DS-source and multi-source models achieve com-

View Article Online
DOI: 10.1039/D5DD00400D

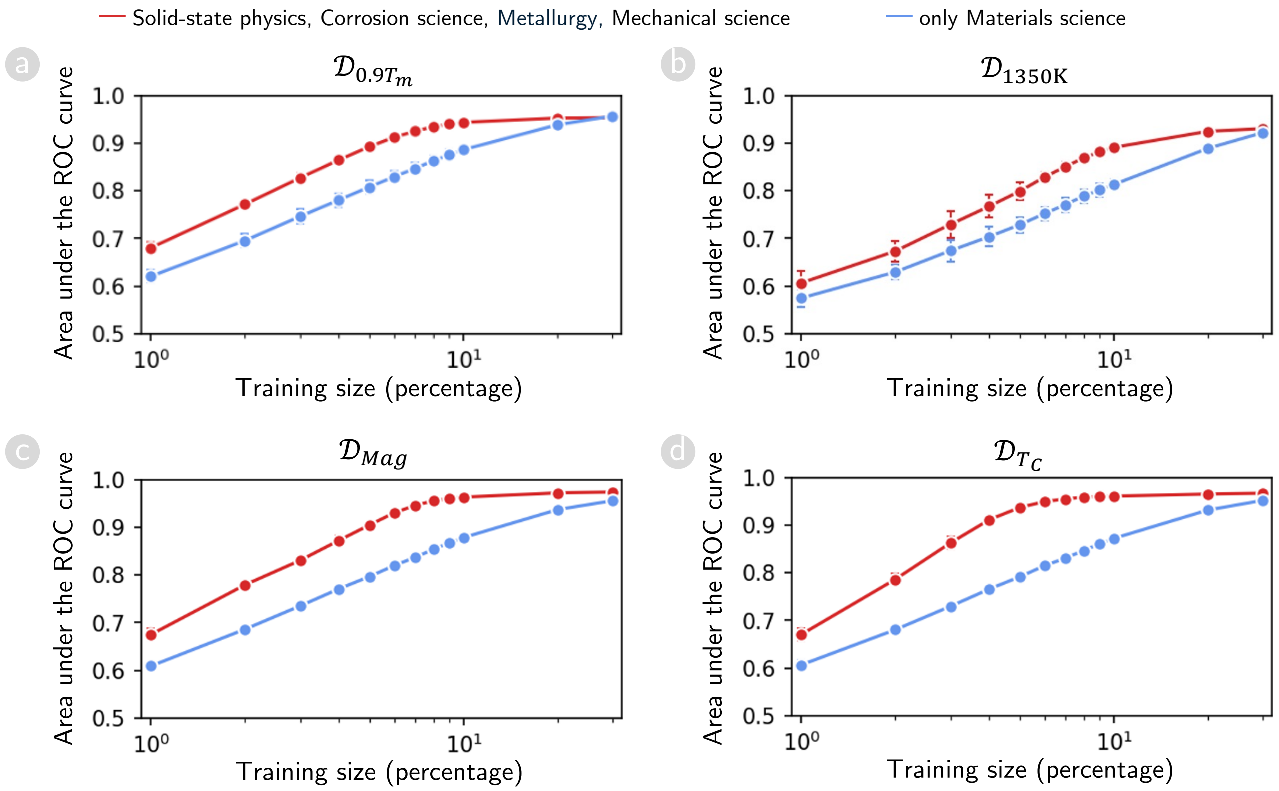


FIG. 6. **Performance comparison of explicit versus implicit domain integration.** Area under ROC curves for predicting HEA stability ($\mathcal{D}_{0.9T_m}$, $\mathcal{D}_{1350\mathrm{K}}$) and magnetic properties ($\mathcal{D}_{Mag}$, $\mathcal{D}_{T_C}$) using two domain integration strategies: (i) systematic combination of four specialized domains (solid-state physics, corrosion science, metallurgy, materials mechanics) shown in red, (ii) only using materials science, which serves as an integrative field that synthesizes perspectives from four specialized domains, shown in blue.

parable AUC early on and remain highly competitive as training data accumulates. For magnetization and Curie-temperature datasets, DS-source models briefly outperform multi-source models at moderate training sizes (approximately 6–20%), but this gap diminishes at larger training sizes.

We note that the LLM-derived substitutability matrix **M** remains fixed across all training sizes (LLMs are used out-of-the-box without retraining); improved performance with larger training sets results from having more host compositions available to apply this fixed knowledge through substitution-based inference (Section II.D). This explains why LLM-source and multi-source models benefit from increased training data despite the LLM knowledge itself remaining unchanged.

Figure 6 provides compelling evidence for the effectiveness of our systematic evidence combination approach compared to relying on materials science as an integrative domain that synthesizes perspectives from the other four domains. Significantly, using only materials science knowledge yields substantially lower performance by 10-20% across all datasets than our multi-source framework, which systematically combines evidence from the four specialized domains, across different prediction tasks. This performance gap demonstrates the fundamental advantage of our Dempster–Shafer-based approach: while materials science provides a static, pre-integrated perspective that may obscure domain-specific nuances, our framework preserves distinct domain insights and adaptively weights them based on their alignment with target properties. The superior performance of our systematic combination method validates that explicit, property-aware evidence synthesis outperforms implicit knowledge fusion, particularly when different domains contribute varying degrees of relevant information for specific material properties such as stability, magnetization, or Curie temperature.

While LLM-source models generally perform well, our results reveal two scenarios where they potentially underperform compared to data-driven approaches.

1. *Property-specific predictions with weak domain alignment:* For magnetic property datasets ($\mathcal{D}_{\mathrm{Mag}}$, $\mathcal{D}_{T_C}$), DS-source substantially outperforms LLM-source, showing a larger performance gap than observed for phase stability datasets (Figures 4 and 5). The five selected domains (corrosion

View Article Online
DOI: 10.1039/D5DD00400D

TABLE IV. Prediction accuracy and Areas under the receiver operating characteristic (ROC) curves of various methods on quaternary-alloy datasets in extrapolation experiments. For each dataset, alloys containing a specific element e are systematically excluded from the training set and used exclusively for testing. Results are reported as mean accuracy and mean AUC, averaged across all elements e within each dataset, with standard deviations reflecting variability across elements.

Evaluation criteria	**Methods**	$\mathcal{D}_{0.9T_m}$	$\mathcal{D}_{1350\text{K}}$	$\mathcal{D}_{\text{Mag}}$	$\mathcal{D}_{T_C}$
Prediction accuracy	Multi-source model	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.86 \pm 0.19}$	$\mathbf{0.86 \pm 0.18}$
	LLM-source model	0.84 ± 0.09	0.90 ± 0.09	0.81 ± 0.21	0.86 ± 0.18
	DS-source model	0.50 ± 0.04	0.51 ± 0.05	0.48 ± 0.07	0.50 ± 0.10
	LR-based model	0.83 ± 0.05	0.91 ± 0.04	0.67 ± 0.15	0.68 ± 0.13
Area under ROC curves	Multi-source model	$\mathbf{0.93 \pm 0.06}$	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.95 \pm 0.06}$	$\mathbf{0.94 \pm 0.07}$
	LLM-source model	0.91 ± 0.11	0.90 ± 0.12	0.95 ± 0.06	0.94 ± 0.07
	DS-source model	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
	LR-based model	0.85 ± 0.11	0.82 ± 0.10	0.84 ± 0.06	0.84 ± 0.06

science, materials mechanics, metallurgy, solid-state physics, materials science) were optimized for structural stability and do not adequately capture magnetic exchange interactions or spin configurations.

2. *Data-rich regimes:* At large training sizes (>20%, Figures 4 and 5), DS-source matches or exceeds LLM-source performance across all datasets. When sufficient data exists, empirical patterns extracted directly from the dataset provide adequate information, and general domain knowledge offers minimal additional value.

In conclusion, LLM-source models excel in data-scarce scenarios by leveraging domain-specific insights to mitigate sparsity-related challenges. As data availability increases, DS-source models outperform LLM-source models, particularly where DS-derived evidence provides sufficient information for a purely data-driven learning approach. Multi-source models, which integrate insights derived from LLM and DS-sources, demonstrate robust and consistent performance across various training sizes.

C. Extrapolation Analysis of Multi-Source Knowledge Integration

Having assessed the proposed framework via cross-validation (Section IV B), we examine its *extrapolation* performance on quaternary alloys containing an element e, which is excluded during training. Unlike the cross-validation experiments, the training set size is not varied for this set of experiments. Instead, for each element e, we remove all e-containing alloys from the dataset and train each model on the remaining alloys that do not contain e. Further, we evaluate the ability of the models to predict the properties of e-containing alloys. This procedure tests whether the learned models can generalize to compositions containing unseen elements in their training datasets.

Table IV reveals distinct performance patterns across model types. DS-source models fail in this scenario, achieving ∼0.50 accuracy (random guessing) across all datasets because they cannot extract substitutability patterns for absent element e from training data. In contrast, LLM-source models achieve substantially higher accuracies across all datasets. Multi-source models modestly outperform LLM-source on phase stability datasets ($\mathcal{D}_{0.9T_m}$ and $\mathcal{D}_{1350\text{K}}$) but achieve nearly identical performance on magnetic property datasets ($\mathcal{D}_{\text{Mag}}$ and $\mathcal{D}_{T_C}$).

This convergence of multi-source and LLM-source performance on magnetic datasets reflects proper uncertainty handling rather than a limitation. When element e is absent from training, DS-source has no observed substitutability patterns involving e. Following the principle established in Section II A, DS-source assigns unit mass to the uncertainty set, explicitly representing total ignorance about e-containing compositions. When this total uncertainty combines with confident LLM evidence through Dempster's rule (Equation 7), the final multi-source prediction is naturally dominated by informative LLM knowledge. The framework thus explicitly represents *unknown* rather than forcing unreliable predictions from insufficient data, demonstrating principled uncertainty quantification in extrapolation scenarios.

Figure 7 illustrates the ROC curves, showing that the multi-source and LLM-source models consistently exhibit higher TPR at comparable FPR across all datasets. Conversely, DS-source models exhibit near-random discrimination, as evidenced by their diagonal ROC curves, while LR-based models yield moderate performance between these extremes. To quantify these visual differences, Table IV also lists AUC for each dataset. Multi-source models achieve the highest AUC scores (0.92–0.95), followed closely by LLM-source models (0.90–0.95), while LR-based models peak at approximately 0.85, and DS-source models hover at approximately 0.50.

Figure 8a–c illustrates knowledge integration in extrapolation simulations for Os-based alloys using the $\mathcal{D}_{0.9T_m}$ dataset. Specifically, Figures 8a and 8b present maps reconstructed from element substitutability patterns derived from the DS-source and multi-source models, respectively, both trained on $\mathcal{D}_{0.9T_m}$ dataset excluding Os-based alloys. Details of the visualization method

View Article Online
DOI: 10.1039/D5DD00400D

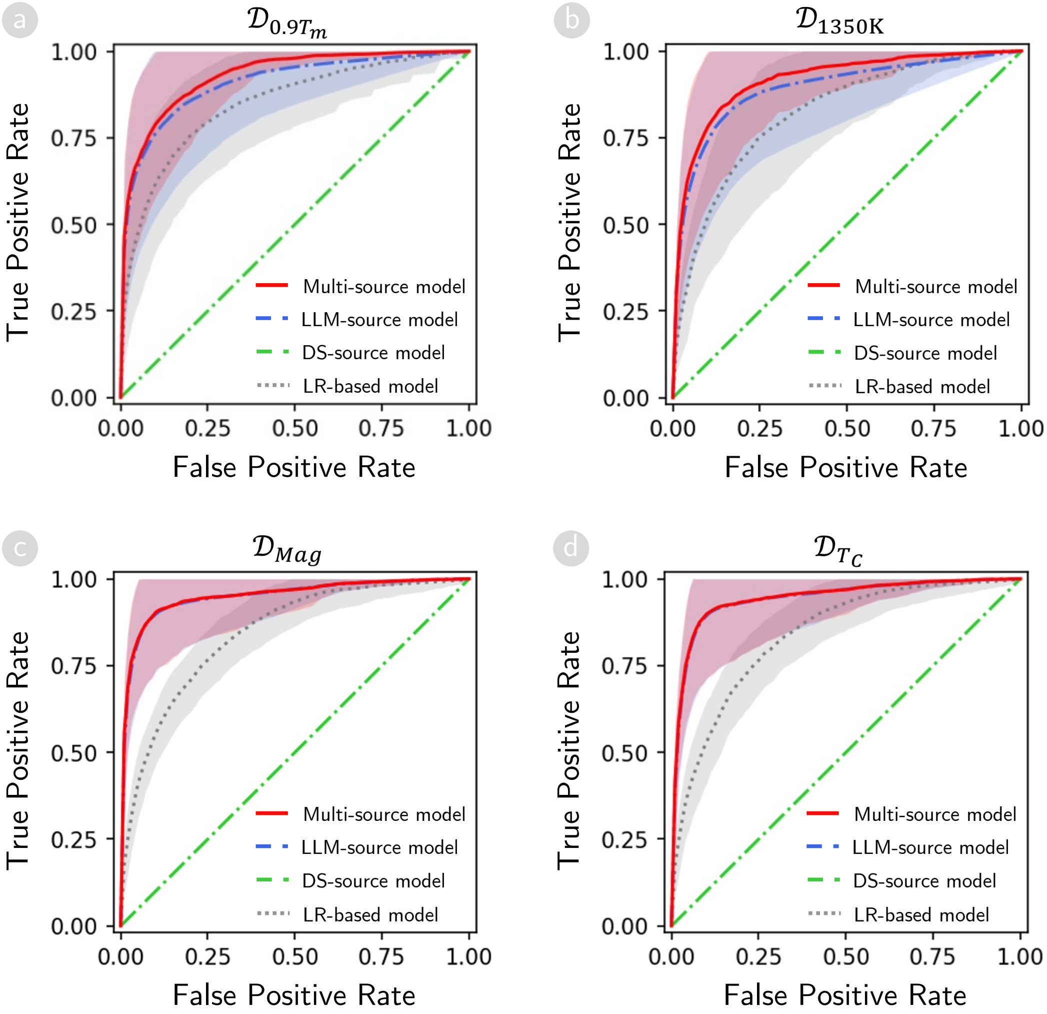


FIG. 7. **Predictive capability evaluation via extrapolation on quaternary-alloy datasets**. For each dataset, alloys containing a specific element e are systematically excluded from the training set and used exclusively for testing. (a–d) Area under the receiver operating characteristic (ROC) curves (AUC) is plotted for each model on their respective test sets in the extrapolation experiments. In all subplots, red lines represent the multi-source model (integrating both DS and LLM sources), green and blue lines represent single-source models (using either DS or LLM sources), and gray lines represent the LR-based model.

are shown in Supplementary Section 4. In these visualizations, the observed alloys are well-structured into sub-clusters according to their phase formation behavior, with blue markers indicating HEA-forming alloys and red markers representing non-HEA alloys. The Os-based candidate alloys, depicted as white circular markers, consistently form a distinct sub-cluster in the upper region of each map. In these visualizations, the background coloration indicates the predicted probability of HEA formation, with deeper blue regions suggesting higher probability of forming stable HEAs.

The limitations of the DS-source model become evident in Figure 8a, where the phase behavior of Os-based alloys remains undetermined due to the absence of Os-containing alloys in the training dataset. This knowledge gap leaves researchers with no guidance when exploring the uncharted territory of Os-based alloys, forcing them to rely on random selection. In contrast, our multi-source approach addresses this limitation by integrating expert insights distilled from scientific literature using LLMs, as illustrated in Figure 8b. The effectiveness of this approach is visually confirmed in Figure 8c, where the multi-source model's predictions closely align with the actual phase behavior of the candidates. This qualitative assessment is complemented by quantitative evaluation in Supplementary Table 4, which reports that the multi-source model achieves an impressive 88% prediction accuracy for Os-based alloys, validating our approach's ca-

View Article Online
DOI: 10.1039/D5DD00400D

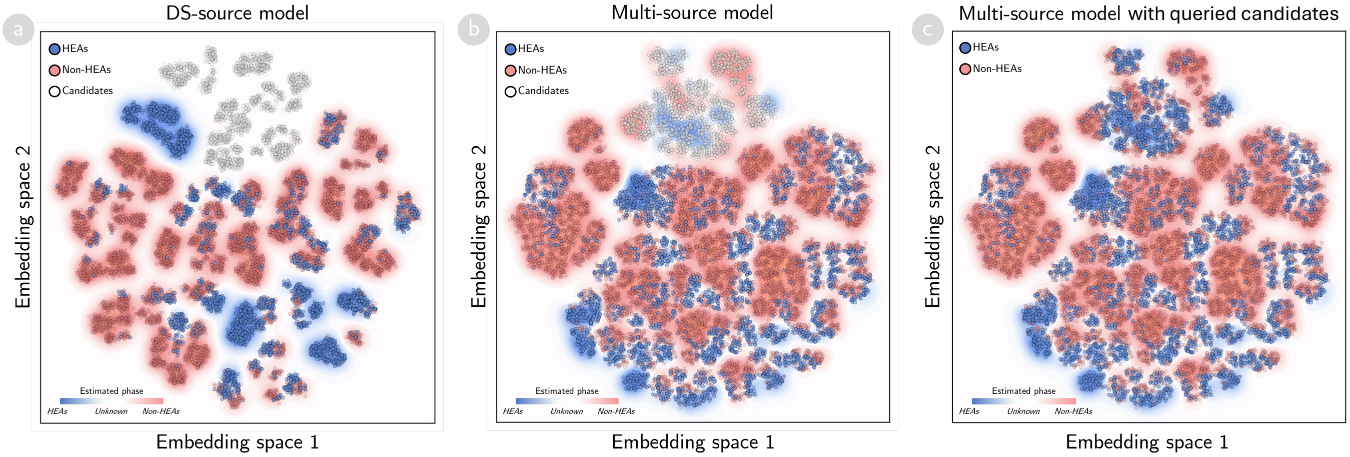


FIG. 8. **Visualization of Os-based alloy extrapolation in dataset $\mathcal{D}_{0.9T_m}$.** (a) Alloy map generated from element substitutability patterns extracted using the DS-source model after excluding Os-based alloys from training. (b–c) Alloy maps generated from element substitutability patterns extracted using the multi-source model after excluding Os-based alloys from training. The map in (c) incorporates queried labels for Os-based candidate alloys. Marker colors represent phase formation: blue for HEA alloys, red for non-HEA alloys, and white for Os-based candidate alloys. Background coloration indicates the predicted phase formation probability according to the DS-source model (a) and multi-source model (b–c), with deeper blue shades suggesting higher probability of HEA formation.

pability to effectively extrapolate to unexplored compositional spaces. In summary, these results confirm that leveraging multi-source or LLM-based evidence significantly enhances discriminative power in the extrapolation scenario.

D. Effectiveness Assessment on Experimental High-Entropy Alloy Data

To assess the real-world applicability of our framework, we validated its performance on experimentally verified alloys from the literature. This validation examines whether the proposed framework, developed primarily using computational datasets, can accurately predict phase stability for experimentally synthesized alloys. Our framework integrates LLM-derived knowledge with substitutability patterns extracted from computational databases using the methodology described in Section II A. This reflects real-world scenarios where researchers must consider all available knowledge before selecting candidates for expensive synthesis.

We performed *5*-fold cross-validation on experimental datasets: $\mathcal{D}^{\text{exp}}_{\text{HEA}}$ of 55 experimentally confirmed alloys. For the HEA dataset $\mathcal{D}^{\text{exp}}_{\text{HEA}}$, we integrated LLM knowledge with substitutability patterns extracted from computational datasets $\mathcal{D}_{\text{1350K}}$, $\mathcal{D}_{\text{AFLOW}}$, $\mathcal{D}_{\text{CALPHAD}}$, and $\mathcal{D}_{\text{LTVC}}$. Details of the computational datasets are introduced in the Supplementary Section 6. Notably, the predictions from these computational methods for the 55 experimentally confirmed alloys are not utilized in our framework training, ensuring unbiased validation.

For benchmarking on the HEA dataset, we compared our framework against four empirical rules (ERs)[58–61], two free-energy models (FEM)[3,62], and a valence-electron concentration (VEC) model[63]. Supplementary Table 2 provides details of these baseline models. Additionally, we compared our framework with results obtained from computational datasets $\mathcal{D}_{\text{AFLOW}}$[15], $\mathcal{D}_{\text{LTVC}}$[19], and $\mathcal{D}_{\text{1350K}}$[45]. These computational datasets are collected by using high-throughput approaches and Hamiltonian models.

Figure 9a presents ROC curves demonstrating that our multi-source integration framework consistently outperforms empirical phase selection models such as ERs, FEMs, and VEC, while achieving performance comparable to costly computational methods. These results confirm that systematically integrating diverse evidence sources through our DST framework enhances prediction accuracy across different material classes. The framework's value does not lie in replacing established methods but in effectively combining their complementary strengths, creating a unified platform that enhances practical decision-making in materials discovery.

To investigate the underlying mechanisms of forming HEAs, we analyzed the elemental substitutability patterns extracted by our framework from multiple evidence sources. Specifically, we integrated substitutability information from the experimental dataset $\mathcal{D}^{\text{exp}}_{\text{HEA}}$, computational datasets ($\mathcal{D}_{\text{1350K}}$, $\mathcal{D}_{\text{AFLOW}}$, $\mathcal{D}_{\text{CALPHAD}}$, $\mathcal{D}_{\text{LTVC}}$), and LLM-derived knowledge.

Figure 9b presents the substitutability matrix for 26 elements relevant to HEA stability, along with their hierarchical clustering structure. The dendrogram is generated via hierarchical agglomerative clustering (HAC) with the complete linkage criterion, grouping elements

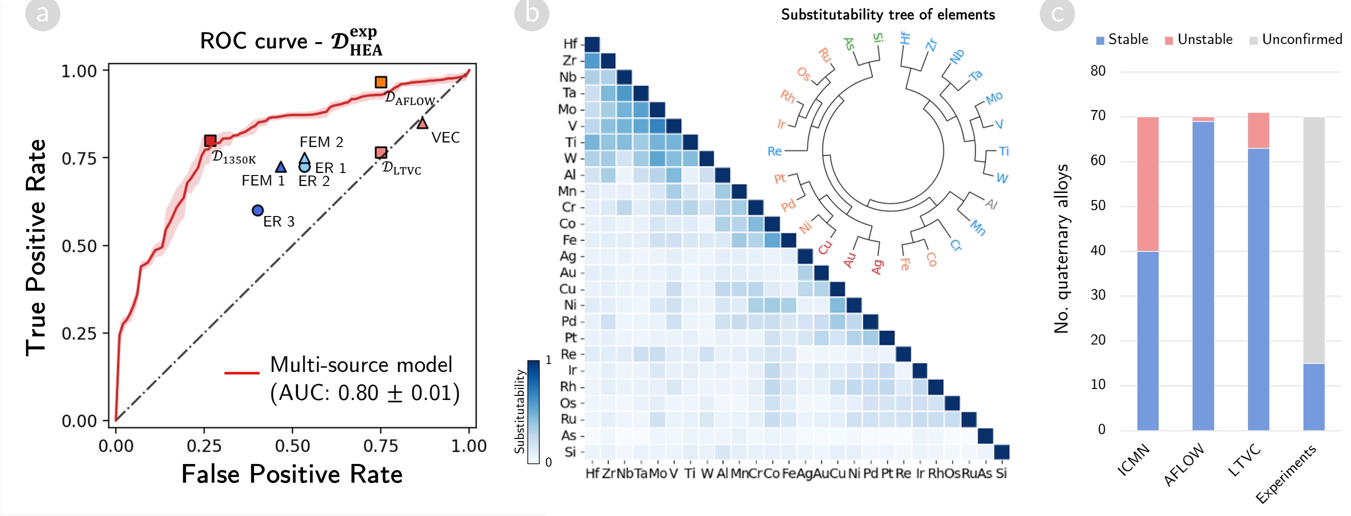


FIG. 9. **Effectiveness Assessment of Multi-Source Knowledge Integration for High-Entropy Alloy Formation**. (a) Receiver operating characteristic (ROC) curves for the phase estimation task on experimental dataset $\mathcal{D}_{\mathrm{HEA}}^{\mathrm{exp}}$. Red line represent the multi-source model (integrating both DS and LLM sources) and gray dashed line represent the random selection. Coloured scatter points represent results of ERs, FEMs, VEC, and computational methods that return only a single stable/unstable estimation. (b) Substitutability matrix and substitutability tree for 26 elements. Matrix values represent substitutability scores derived from integrated computational datasets, experimental dataset and LLM sources. The substitutability tree is generated using hierarchical agglomerative clustering with complete linkage criterion. Element colors: blue (early transition metals), orange (intermediate transition metals), gray (post-transition elements). (c) Predicted phase stability for 70 possible quaternary alloys from Group 1 elements (Hf, Zr, Nb, Ta, Mo, V, Ti, W). Bars show number of alloys predicted as single-phase obtained from computational datasets ($\mathcal{D}_{\mathrm{AFLOW}}$[15], $\mathcal{D}_{\mathrm{LTVC}}$[19], and $\mathcal{D}_{1350\mathrm{K}}$[45]) and experimentally verified single-phase HEAs[45,49,50].

based on similar substitutability patterns. The substitutability analysis reveals three distinct element groups with strong intra-group substitutability. Group 1 comprises eight early transition metals from periodic groups 4–6: Ti, Zr, Hf (group 4); V, Nb, Ta (group 5); and Mo, W (group 6). Cr, while belonging to group 6, exhibits unique behavior, showing moderate substitutability with Group 1 elements but high substitutability with Fe, Co, Mn, and Al, which together form Group 2. Group 3 contains primarily late transition metals from periodic groups 9–11, including Rh, Ir, Pd, Pt, Ni, Cu, Au, Ag. Notably, Groups 1 and 3 show weak inter-group substitutability but moderate substitutability with the bridging Group 2.

The exceptional intra-group substitutability of Group 1 elements (Ti, Zr, Hf, V, Nb, Ta, Mo, W), exhibiting notably higher scores than Groups 2 and 3, suggests a design principle: quaternary combinations should readily form stable single-phase HEAs. Critically, this substitutability matrix (Figure 9b) is derived by fusing evidence from multiple independent sources–experimental HEA dataset ($\mathcal{D}_{\mathrm{HEA}}^{\mathrm{exp}}$), computational databases ($\mathcal{D}_{1350\mathrm{K}}$, $\mathcal{D}_{\mathrm{AFLOW}}$, $\mathcal{D}_{\mathrm{LTVC}}$), and 20 LLM-domain sources–through Dempster–Shafer integration; such high mutual substitutability indicates unanimous agreement across all sources regarding these patterns. Figure 9c validates this prediction: all three computational datasets unanimously predict single-phase formation for all 70 possible Group 1 quaternaries, and all 15 experimentally synthesized compositions form single-phase HEAs (100% success rate). This agreement is consistent with established principles for refractory high-entropy alloys[41,64]: early transition metals (groups 4–6) preferentially form stable BCC solid solutions due to similar atomic sizes and compatible electronic structures, with single-phase stability thermodynamically reinforced by configurational entropy that lowers Gibbs free energy at elevated temperatures[65].

E. Effectiveness Assessment on Experimental High-Entropy Boride Data

We extend our analysis to high-entropy borides (HEBs), where boron's restrictive bonding requirements create similarly high elemental selectivity as observed in HEAs[66]. Despite different underlying mechanisms, both systems share the key challenge of identifying rare viable combinations within vast compositional spaces, making HEBs suitable for demonstrating our framework's applicability to diverse multi-component materials with stringent compatibility constraints.

In this experiment, we applied our framework to a dataset of 19 experimentally confirmed quinary borides collected from previous studies. Using these validated compositions as training data, our framework was then employed to rank 314 potential quinary boride candi-

View Article Online
DOI: 10.1039/D5DD00400D

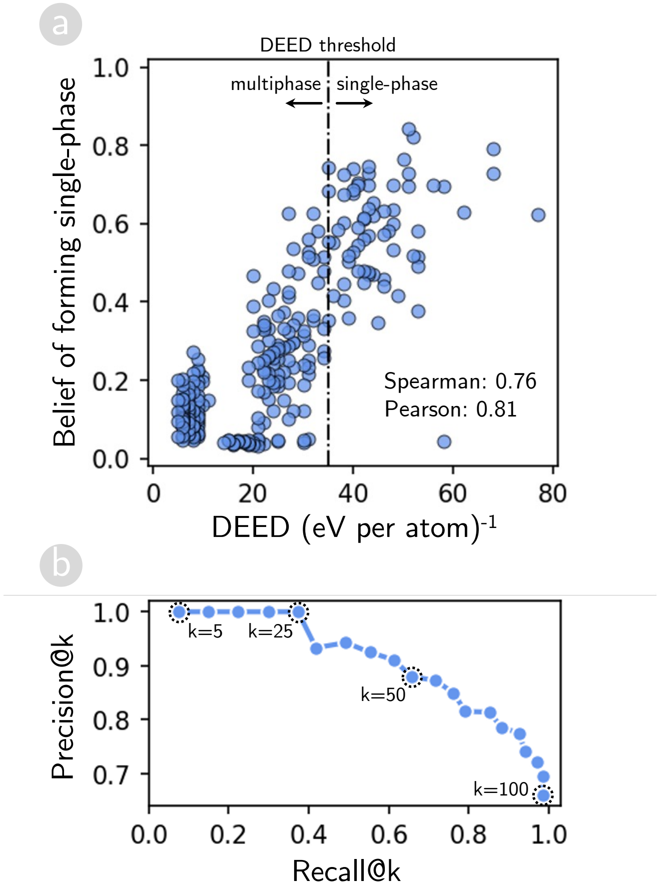


FIG. 10. **Effectiveness Assessment of Multi-Source Knowledge Integration for High-Entropy Borides Formation**. (a) Correlation analysis between our framework's single-phase formation belief and the disordered enthalpy-entropy descriptors (DEED) for 275 quinary boride candidates. The dashed line indicates the DEED threshold of 35 (eV per atom)$^{-1}$ for single-phase prediction. (b) Precision@k and Recall@k performance metrics evaluated at k values from 5 to 100 with increments of 5

dates formed by boron as the anion and the following metals: Cr, Hf, Ir, Mn, Mo, Nb, Ta, Ti, V, W, Y, Zr. To benchmark our framework, we compared the rankings obtained by our framework with those derived using the disordered enthalpy-entropy descriptors (DEED)[44], which represents the state-of-the-art descriptor based on ab-initio calculations for guiding experimental discovery of new single-phase high-entropy carbonitrides and borides.

Figure 10a illustrates the correlation between DEED values and the belief of forming single-phase structures for 275 of the 314 quinary boride candidates. For the remaining 39 candidates, our framework could not provide reliable predictions due to insufficient training data coverage, resulting in maximum uncertainty values that rendered these predictions uninformative for comparison purposes. The results demonstrate a strong positive linear correlation between the single-phase formation belief derived from our framework and the DEED values, with Pearson and Spearman correlation coefficients of 0.81 and 0.76, respectively. The previous DEED study established a threshold of 35 (eV per atom)$^{-1}$ to distinguish between single-phase and multiphase candidates, where values above this threshold indicate predicted single-phase formation.

The strong correlation for the 275 confident predictions, combined with explicit uncertainty flagging for 39 candidates, demonstrates effective uncertainty quantification. To further validate this mechanism, we analyzed prediction accuracy at varying uncertainty thresholds, as shown in Supplementary Figure 8. The results reveal a systematic trade-off: as the uncertainty threshold decreases (accepting more uncertain predictions as confident), prediction accuracy degrades accordingly. This behavior confirms that high uncertainty values successfully flag regions where evidence is insufficient, preventing overconfident extrapolation beyond the training data. The explicit uncertainty quantification thus serves as a critical safeguard against overfitting in data-sparse scenarios, distinguishing our approach from conventional machine learning methods that would force predictions regardless of data sufficiency.

To evaluate our framework's practical utility as a materials discovery tool, we analyzed how well it ranks promising candidates compared to the established DEED method. We measured this using standard ranking metrics: Precision@k (what percentage of our top k recommendations are actually good) and Recall@k (what percentage of all good candidates we capture in our top k recommendations). The results show impressive performance: when we look at our top 25 recommendations (k=25), all of them were also predicted to form single-phase structures by the DEED method, giving us perfect precision, as shown in Figure 10b. More broadly, to capture 50% of all the promising candidates identified by DEED, our method requires selecting approximately the top 35-40 candidates and maintains over 90% precision, meaning that more than 90% of these top-ranked candidates are correctly identified as single-phase according to DEED. Even when capturing 75% of the promising candidates, our precision remains above 85%. These results demonstrate that our framework effectively prioritizes the most promising compositions for experimental synthesis.

The strong performance on high-entropy borides, combined with the previous results on high-entropy alloys, establishes the framework's capability to handle uncertainty in compositionally selective multi-component material systems. Notably, while computational databases such as AFLOW and CALPHAD carry inherent uncertainties from DFT approximations and thermodynamic extrapolations[18], the Dempster–Shafer theory explicitly models these through mass assignments to ignorance, enabling robust integration with experimental data and

View Article Online
DOI: 10.1039/D5DD00400D

mitigating risks of systematic errors in guiding alloy synthesis. The discount factor mechanism (Equations 5–7) automatically downweights unreliable sources based on cross-validation performance, preventing error propagation by allowing high-quality evidence to dominate when computational predictions conflict with experimental observations.

F. Limitations and Future Extensions

Previous sections have demonstrated the framework's effectiveness across computational and experimental datasets. We now examine its current limitations and corresponding opportunities for future development.

Context-Independent Evidence Weighting: The current implementation employs fixed weighting parameters for each source without considering the specific context of elemental substitution. For instance, metallurgical knowledge may be more reliable for refractory elements, while solid-state physics insights may better inform noble metal substitutability. Future extensions could implement context-dependent weighting, wherein discount factors vary based on the element pair under consideration. This could be achieved by conditioning discount factors on elemental properties such as atomic radius, electronegativity, or periodic group membership, enabling the framework to recognize element-specific reliability patterns across different knowledge sources.

From Uncertainty Quantification to Discovery Navigation: This study proposes a framework to integrate multi-source knowledge and quantify uncertainty for candidate materials. However, a subsequent challenge remains: how to effectively utilize these uncertainty measures to select candidates for experimental validation under limited resources. This candidate selection problem inherently involves balancing exploration (investigating compositions with high uncertainty that may reveal novel alloys) and exploitation (refining predictions in promising regions with moderate uncertainty). Active learning provides a principled approach to this challenge by identifying experiments that maximally reduce epistemic uncertainty, prioritizing candidates where additional data would most improve model reliability. Reinforcement learning complements this by learning optimal selection policies through iterative experimental feedback, dynamically adjusting the exploration–exploitation balance as the discovery campaign progresses. Together, these techniques could transform the current prediction framework into a comprehensive decision-support system for accelerated materials discovery.

Symmetric Substitutability Assumption: The symmetric substitutability assumption (A$\rightarrow$B and B$\leftarrow$A are equivalent) represents a context-averaged approximation that may limit accuracy for systems with strong directional substitution preferences. This symmetric treatment is justified in this study by two factors: first, the limited training data in our data-sparse scenarios makes learning separate directional patterns statistically infeasible; second, for near-equiatomic multi-principal element HEAs characterized by disordered random solid solutions, elements occupy statistically similar local environments, rendering symmetric substitution a physically reasonable first-order approximation. Future extensions could incorporate asymmetric substitutability by maintaining separate A$\rightarrow$B and B$\leftarrow$A matrices and collecting directional evidence from LLMs through modified prompts.

Broaden Scope Beyond Phase Stability: To serve the purpose of screening the element combinations forming HEA phases, the proposed framework focuses on the fundamental question of whether the HEA phase exists. We design a frame of discernment $\Omega_{HEA} = \{\text{HEA}, \overline{\text{HEA}}\}$ to model the existence of HEA phases with mass functions. Consequently, our framework has not answered essential questions regarding the structure and other properties of the HEAs. However, by redesigning the frame of discernment to reflect the additional properties of interest, we can also construct a model that can recommend potential alloys forming HEA phases with desirable properties. Extending to mechanical, electronic, or catalytic properties represents another promising direction as sufficient property-specific data becomes available[67].

Scalability to Higher-Order Systems: The current validation focuses primarily on quaternary alloy systems, with limited exploration of higher-order compositions. Extension to quinary and higher-order alloys could be achieved through hierarchical decomposition, wherein quaternary systems serve as baseline evidence augmented by pairwise substitutability relationships. However, more complex systems may require sparse approximation techniques and substantially larger materials databases to maintain predictive reliability.

V. CONCLUSIONS

The central contribution of this work lies in demonstrating that the interpolation–extrapolation dichotomy inherent to conventional data-driven materials discovery can be systematically addressed through principled integration of multi-source knowledge. Crucially, the framework does not indiscriminately combine all available evidence; rather, it evaluates the reliability of each source based on its alignment with the target property, ensuring that only relevant domain knowledge contributes meaningfully to predictions. By employing elemental substitutability as a unifying concept and leveraging Dempster–Shafer theory to combine empirical observations with insights extracted from scientific literature via LLMs, the framework effectively bridges data-rich and data-sparse regions in materials exploration. Our framework demonstrates superior performance compared to traditional data-driven approaches and empirical phase selection rules, while achieving accuracy comparable to computationally expensive methods, particularly when predicting phase stability for compositions containing previously unseen elements. These results highlight that the significance of the framework does not reside in superseding established methods, but rather in effectively synthesizing their complementary strengths while representing epistemic limitations transparently.

Beyond HEAs, this framework could accelerate discovery in several materials classes facing similar challenges of vast compositional spaces and sparse data, including functional ceramics[44], and catalytic materials[34]. Through successful validation on diverse alloy systems, this study demonstrates that uncertainty-aware AI integration provides a viable path forward for accelerated materials discovery. The element substitutability patterns extracted using this framework may also inform synthetic strategies for targeted property optimization across diverse material applications.

AUTHOR CONTRIBUTIONS

M.-Q. H.: Conceptualization, Methodology, Software, Formal analysis, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. D.-K. L.: Software, Investigation, Data Curation. V.-C. N.: Software, Formal analysis, Data Curation. H. K.: Investigation, Validation, Writing - Review & Editing. S. C.: Investigation, Validation, Writing - Review & Editing. H.-C. D.: Conceptualization, Methodology, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

CONFLICT OF INTERESTS

The authors report there are no competing interests to declare.

DATA AVAILABILITY

Publicly Available Datasets: Data for this article, including experimental and computational datasets supporting high-entropy alloy phase prediction, are available at Zenodo at `https://doi.org/10.5281/zenodo.17074832`.

Code Availability: Code for the uncertainty-aware AI integration framework is available at GitHub at `https://github.com/minhquyet2308/Uncertainty-Aware-AI-Intergration`, with an archived version available at Zenodo at `https://doi.org/10.5281/zenodo.17744151`.

ACKNOWLEDGEMENTS

This work is supported by the JST-CREST Program (Innovative Measurement and Analysis), under Grant number JPMJCR2235; and the JSPS KAKENHI Grant Numbers 20K05301, JP19H05815, 20K05068, 23KJ1035, 23K03950, and JP23H05403. S.C. acknowledges support by US-DoD (ONR MURI program N00014-21-1-251). H. K. acknowledges support by the Japan Science and Technology Agency (JST) ASPIRE Program under the project "International Collaborative Research Network for Advanced Atomic Layer Processes". The authors thank Dr. Huan Tran and Dr. Xiomara Campilongo for fruitful discussions.

REFERENCES

[1] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, and S.-Y. Chang, Advanced Engineering Materials **6**, 299 (2004).

[2] B. Cantor, I. Chang, P. Knight, and A. Vincent, Materials Science and Engineering: A **375-377**, 213 (2004).

[3] O. Senkov and D. Miracle, Journal of Alloys and Compounds **658** (2015), 10.1016/j.jallcom.2015.10.279.

[4] J. M. Rickman, H. M. Chan, M. P. Harmer, J. A. Smeltzer, C. J. Marvel, A. Roy, and G. Balasubramanian, Nature Communications **10**, 2618 (2019).

[5] M.-H. Tsai and J.-W. Yeh, Materials Research Letters **2**, 107 (2014).

[6] C. Toher, C. Oses, D. Hicks, and S. Curtarolo, npj Comput. Mater. **5**, 69 (2019).

[7] G. Deshmukh, N. J. Wichrowski, N. Evangelou, P. G. Ghanekar, S. Deshpande, I. G. Kevrekidis, and J. Greeley, npj Computational Materials **10**, 116 (2024).

[8] M. Ghorbani, M. Boley, P. N. H. Nakashima, and N. Birbilis, Scientific Reports **14**, 8299 (2024).

[9] M.-H. Tsai, Entropy **18**, 252 (2016).

[10] M.-H. Tsai, R.-C. Tsai, T. Chang, and W.-F. Huang, Metals **9**, 247 (2019).

[11] W. Huang, P. Martin, and H. L. Zhuang, Acta Mater. **169**, 225 (2019).

[12] Z. Rao, P.-Y. Tung, R. Xie, Y. Wei, H. Zhang, A. Ferrari, T. Klaver, F. Körmann, P. T. Sukumar, A. K. da Silva, Y. Chen, Z. Li, D. Ponge, J. Neugebauer, O. Gutfleisch, S. Bauer, and D. Raabe, Science **378**, 78 (2022).

[13] J. Roberts, B. Rijal, S. Divilov, J.-P. Maria, W. G. Fahrenholtz, D. E. Wolfe, D. W. Brenner, S. Curtarolo, and E. Zurek, npj Computational Materials **10**, 142 (2024).

[14] D. Alman, Entropy **15**, 4504 (2013).

[15] F. Zhang, C. Zhang, S. Chen, J. Zhu, W. Cao, and U. Kattner, Calphad **45**, 1 (2014).

[16] M. Esters, C. Oses, S. Divilov, H. Eckert, R. Friedrich, D. Hicks, M. J. Mehl, F. Rose, A. Smolyanyuk, A. Calzolari, X. Campilongo, C. Toher, and S. Curtarolo, Comp. Mat. Sci. **216**, 111808 (2023).

[17] C. Oses, M. Esters, D. Hicks, S. Divilov, H. Eckert, R. Friedrich, M. J. Mehl, A. Smolyanyuk, X. Campilongo, A. van de Walle, J. Schroers, A. G. Kusne, I. Takeuchi, E. Zurek, M. Buongiorno Nardelli, M. Fornari, Y. Lederer, O. Levy, C. Toher, and S. Curtarolo, Comp. Mat. Sci. **217**, 111889 (2023).

[18] C. Toher and S. Curtarolo, Journal of Phase Equilibria and Diffusion **45**, 219 (2024).

[19] Y. Lederer, C. Toher, K. S. Vecchio, and S. Curtarolo, Acta Materialia **159**, 364 (2018).

[20] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, npj Comput. Mater. **4**, 29 (2018).

[21] G. L. W. Hart, T. Mueller, C. Toher, and S. Curtarolo, Nature Reviews Materials **6**, 730 (2021).

[22] M.-Q. Ha, D.-N. Nguyen, V.-C. Nguyen, T. Nagata, T. Chikyow, H. Kino, T. Miyake, T. Denœux, V.-N. Huynh, and H.-C. Dam, Nature Computational Science **1**, 470 (2021).

[23] J. He, R. Yin, C. Wang, C. Liu, D. Xue, Y. Su, L. Qiao, T. Lookman, and Y. Bai, Journal of Materiomics **11**, 100913 (2025).

[24] T. L. Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. C. Dam, Science and Technology of Advanced Materials **18**, 756 (2017), pMID: 29152012.

[25] E. Hüllermeier and W. Waegeman, Machine Learning **110**, 457 (2021).

[26] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," (2010), arXiv:1012.2599 [cs.LG].

[27] J. Snoek, H. Larochelle, and R. P. Adams, in *Advances in Neural Information Processing Systems*, Vol. 25, edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc., 2012).

[28] E. Hüllermeier and K. Brinker, Fuzzy Sets and Systems **159**, 2337 (2008), theme: Information Processing.

[29] E. P. George, D. Raabe, and R. O. Ritchie, Nat. Rev. Mater. **4**, 515 (2019).

[30] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, and A. Maeda, Phys. Rev. B **103**, 014509 (2021).

[31] A. P. Dempster, Journal of the Royal Statistical Society: Series B (Methodological) **30**, 205 (1968).

[32] G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, 1976).

[33] T. Denœux, D. Dubois, and H. Prade, in *A Guided Tour of Artificial Intelligence Research*, Vol. 1, edited by P. Marquis, O. Papini, and H. Prade (Springer Verlag, 2020) Chap. 4, pp. 119–150.

[34] N. Nu Thanh Ton, M.-Q. Ha, T. Ikenaga, A. Thakur, H.-C. Dam, and T. Taniike, 2D Materials **8**, 015019 (2020).

[35] M.-Q. Ha, D.-N. Nguyen, V.-C. Nguyen, H. Kino, Y. Ando, T. Miyake, T. Denœux, V.-N. Huynh, and H.-C. Dam, Journal of Applied Physics **133**, 053904 (2023).

[36] E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, npj Computational Materials **8**, 84 (2022).

[37] D. H. Cook, P. Kumar, M. I. Payne, C. H. Belcher, P. Borges, W. Wang, F. Walsh, Z. Li, A. Devaraj, M. Zhang, M. Asta, A. M. Minor, E. J. Lavernia, D. Apelian, and R. O. Ritchie, Science **384**, 178 (2024).

[38] S. Liu, T. Wen, A. S. Pattamatta, and D. J. Srolovitz, Materials Today **80**, 240 (2024).

[39] Z. Chen, Y. Liu, and H. Sun, Nature Communications **12**, 6136 (2021).

[40] B. Cantor, K. Kim, and P. J. Warren, in *Metastable, Mechanically Alloyed and Nanocrystalline Materials (2001)*, Journal of Metastable and Nanocrystalline Materials, Vol. 13 (Trans Tech Publications Ltd, 2002) pp. 27–32.

[41] D. Miracle and O. Senkov, Acta Materialia **122**, 448 (2017).

[42] A. Tversky, Psychological Review **84**, 327 (1977).

[43] P. Smets, International Journal of Approximate Reasoning **9**, 1 (1993).

[44] S. Divilov, H. Eckert, D. Hicks, C. Oses, C. Toher, R. Friedrich, M. Esters, M. J. Mehl, A. C. Zettel, Y. Lederer, E. Zurek, J.-P. Maria, D. W. Brenner, X. Campilongo, S. Filipović, W. G. Fahrenholtz, C. J. Ryan, C. M. DeSalle, R. J. Crealese, D. E. Wolfe, A. Calzolari, and S. Curtarolo, Nature **625**, 66 (2024).

[45] W. Chen, A. Hilhorst, G. Bokas, S. Gorsse, P. J. Jacques, and G. Hautier, Nature Communications **14**, 2856 (2023).

[46] A. Takeuchi and A. Inoue, MATERIALS TRANSACTIONS **46**, 2817 (2005).

[47] A. Takeuchi and A. Inoue, Intermetallics **18**, 1779 (2010).

[48] T. Fukushima, H. Akai, T. Chikyow, and H. Kino, Phys. Rev. Materials **6**, 023802 (2022).

[49] C. K. H. Borg, C. Frey, J. Moh, T. M. Pollock, S. Gorsse, D. B. Miracle, O. N. Senkov, B. Meredig, and J. E. Saal, Scientific Data **7**, 430 (2020).

[50] G. Khanna R, M. K. Singh, D. K. Rai, and S. Samal, Materials Letters **365**, 136404 (2024).

[51] M. P. LaValley, Circulation **117**, 2395 (2008).

[52] A. Seko, A. Togo, and I. Tanaka, "Descriptors for machine learning of materials data," in *Nanoinformatics* (Springer Singapore, Singapore, 2018) pp. 3–23.

[53] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Phys. Rev. B **95**, 144110 (2017).

[54] F. C. T., Nature **138**, 7 (1936).

[55] U. Mizutani, MRS Bulletin **37**, 169 (2012).

[56] H. O. M. S. E. Mueller, in *Alloy Phase Diagrams* (ASM International, 2016).

[57] Z. Pei, J. Yin, P. K. Liaw, and D. Raabe, Nature Communications **14**, 54 (2023).

[58] X. Yang and Y. Zhang, Materials Chemistry and Physics **132**, 233 (2012).

[59] S. Guo, Q. Hu, C. Ng, and C. Liu, Intermetallics **41**, 96 (2013).

[60] W. Zhijun, Y. Huang, Y. Yang, J. Wang, and C. Liu, Scripta Materialia **94** (2015), 10.1016/j.scriptamat.2014.09.010.

[61] A. K. Singh, N. Kumar, A. Dwivedi, and A. Subramaniam, Intermetallics **53**, 112 (2014).

[62] M. C. Troparevsky, J. R. Morris, P. R. C. Kent, A. R. Lupini, and G. M. Stocks, Phys. Rev. X **5**, 011041 (2015).

[63] S. Guo, C. Ng, J. Lu, and C. T. Liu, Journal of Applied Physics **109**, 103505 (2011).

[64] O. Senkov, G. Wilks, D. Miracle, C. Chuang, and P. Liaw, Intermetallics **18**, 1758 (2010).

[65] B. S. Murty, J.-W. Yeh, S. Ranganathan, and P. P. Bhattacharjee, *High-Entropy Alloys*, 2nd ed. (Elsevier, Amsterdam, 2019).

[66] J. Gild, Y. Zhang, T. Harrington, S. Jiang, T. Hu, M. C. Quinn, W. M. Mellor, N. Zhou, K. Vecchio, and J. Luo, Scientific Reports **6**, 37946 (2016).

[67] S. Nakanowatari, K. Takahashi, H. C. Dam, and T. Taniike, ACS Catalysis **15**, 8691 (2025).

View Article Online
DOI: 10.1039/D5DD00400D

Data Availability Statement (DAS)

Publicly Available Datasets: Data for this article, including experimental and computational datasets supporting high-entropy alloy phase prediction, are available at Zenodo at https://doi.org/10.5281/zenodo.17074832 .

Code Availability: Code for the uncertainty-aware AI integration framework is available at GitHub at https://github.com/minhquyet2308/Uncertainty-Aware-AI-Intergration , with an archived version available at Zenodo at https://doi.org/10.5281/zenodo.17744151 .