Data in Brief 65 (2026) 112540

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A dataset of precipitate-containing multi-principal element alloys

Anshu Raj[a], Xin Wang[b], Matthew Luebbe[c], Haiming Wen[c], Kun Lu[b,*], Shuozhi Xu[a,*]

[a] School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, OK 73019, USA
[b] School of Library and Information Studies, University of Alabama, Tuscaloosa, AL 35487, USA
[c] Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

ARTICLE INFO

Article history:
Received 28 December 2025
Revised 19 January 2026
Accepted 26 January 2026
Available online 30 January 2026

Dataset link: Multi-Category Materials Information Extraction (Composition, Processing, Microstructure, Properties) (Zenodo)

Keywords:
Materials science
Composition-processing-microstructure-property
relationship
Precipitate
Multi-principal element alloys

ABSTRACT

We report a curated dataset that brings together composition, processing conditions, microstructural details, and mechanical properties for 396 combinations of alloy composition and processing condition drawn from 100 peer-reviewed research articles on precipitate-containing multi-principal element alloys (MPEAs). The dataset was created by first utilizing a generative large language model for information extraction, followed by expert review to ensure accurate recovery of materials data. Compositional information was taken directly from tables and text, while processing routes — including homogenization, rolling, recrystallization, and aging — were converted into uniform temperature and time metrics. Microstructural descriptors, including precipitate phases and sizes, were consolidated into a consistent labeling scheme to accommodate the wide range of terminology used in published literature. Finally, mechanical property data, such as strength and ductility, were compiled together with the temperatures at which they were measured. These data provide a coherent view of the composition-processing-microstructure-property features explored in existing MPEA research and establish a resource that supports data-driven alloy design as well as future development of automated materials

* Corresponding authors.
E-mail addresses: klu@ua.edu (K. Lu), shuozhixu@ou.edu (S. Xu).
Social media: @RajAnshu009 (A. Raj)

https://doi.org/10.1016/j.dib.2026.112540

information-extraction methodologies. The complete dataset is available on Zenodo.

Specifications Table

Subject	Materials Science: Metals and Alloys
Specific subject area	Multi-principal element alloys
Type of data	Excel (.xlsx)
Data collection	Data, which are on composition-processing-microstructure-property relationships, were collected from 100 peer-reviewed articles on precipitate-containing multi-principal element alloys. A generative large language model-based extraction pipeline was first used for information extraction. Subsequently, the outputs underwent rigorous expert review, wherein researchers cross-referenced the extracted data against the original publications, thereby guaranteeing the fidelity of the final dataset presented here.
Data source location	Institution: University of Oklahoma City: Norman Country: USA Coordinates: 35.1987° N, 97.4449° W
Data accessibility	Repository: Zenodo Identifier: DOI: 10.5281/zenodo.18021833 Direct URL: https://zenodo.org/records/18021833

1. Value of the Data

- These data expand our current knowledge base and understanding of multi-principal element alloys (MPEAs).
- The dataset compiles 396 combinations of alloy composition and processing condition extracted from 100 peer-reviewed publications, encompassing 47 well-defined features that collectively describe composition, processing conditions, microstructural characteristics, and mechanical properties.
- The integrated data make it possible to examine quantitative links between processing conditions, microstructural evolution, and resulting material properties, offering new opportunities for data-driven alloy design and discovery.
- Because every extracted value is directly connected to its original text source, the dataset ensures complete transparency and traceability, allowing researchers to verify, reproduce, and benchmark results with confidence.
- The dataset will benefit researchers in materials science and engineering by providing a reliable benchmark and training resource for developing and testing data-driven models aimed at automated materials information extraction and predictive modeling.

2. Background

MPEAs have been an active area of research for over two decades [1,2]. These alloys often combine the beneficial properties of several principal elements, leading to promising applications in areas such as high-temperature strength, low-temperature ductility, excellent corrosion resistance, and outstanding wear resistance [3,4]. However, the potential composition space of MPEAs is enormous. More than thirty elements can be combined in various ratios, creating mil-

lions of possible compositions [5]. Navigating this vast design space effectively requires data-driven strategies that can link composition, processing, microstructure, and property.

Data-driven approaches, as the name suggests, require a large amount of high-quality data. Although computational materials databases have grown rapidly in recent years, most experimentally derived information, such as processing conditions, microstructural details, and mechanical properties, remains embedded within unstructured scientific text [6,7]. The lack of accessible experimental data limits the ability to train and validate data-driven models for predicting material behavior and designing new alloys.

To address this challenge, the accompanying research developed a multi-stage extraction framework powered by generative large language models (LLMs) to automatically collect, organize, and validate materials information from published literature [8]. The extracted data were subsequently reviewed and revised by experts. The dataset focuses on precipitate-containing MPEAs, which display complex microstructures and diverse mechanical responses [9,10]. By extracting 47 features from 100 peer-reviewed articles that cover composition, processing, microstructure, and properties, this work provides a consistent and source-tracked dataset that connects unstructured text with structured data resources and establishes a reliable foundation for materials informatics.

3. Data Description

The dataset is provided in the xlsx format and made available through the Zenodo platform [11]. It contains 47 columns of information. In total, the file includes 398 rows, representing 100 journal articles that collectively covering 396 combinations of alloy composition and processing condition. The first two rows serve as headers for the extracted information.

- Columns A and B contain the names of the journal articles and the corresponding materials discussed in each article.
- Columns C through P provide the elemental compositions of the materials. The elements included are Fe, Ni, Co, Mn, Cr, Al, Ti, Cu, Si, V, Nb, B, Mo, and Ta.
- Columns Q through AB describe the processing methods applied to the materials.
  - Column Q indicates whether homogenization was performed.
  - Column R identifies the homogenization temperature in °C.
  - Column S specifies the homogenization time in hours.
  - Column T indicates whether rolling was carried out.
  - Column U lists the rolling temperature in °C.
  - Column V shows the rolling percentage.
  - Column W indicates whether recrystallization was performed.
  - Column X lists the recrystallization temperature in °C.
  - Column Y specifies the recrystallization time in minutes.
  - Column Z indicates whether aging was performed.
  - Column AA lists the aging temperature in °C.
  - Column AB specifies the aging time in hours.
- Columns AC through AM summarize the matrix and precipitate phases present in the materials.
  - Column AC identifies the matrix phase used in the material and assigns numerical labels to each phase: face-centered cubic (FCC) = 1, body-centered cubic (BCC) = 2, $L1_2$ = 3, B2 = 4, sigma (σ) = 5).
  - Column AD records the volume percentage of the matrix phase.
  - Column AE identifies the first precipitate type found in the matrix and assigns numerical labels: None = 0, $L1_2$ = 1, gamma double prime (γ'') = 2, B2 = 3, eta (η) = 4, $L2_1$ = 5, sigma (σ) = 6, BCC = 7, FCC = 8, mu (μ) = 9, hexagonal-close packed (HCP) = 10, Laves = 11, epsilon (ε) = 12, γ' = 13, $D0_{19}$ = 14, A2 = 15, and delta (δ) = 16.
  - Column AF notes the size of the first precipitate type in nm.

- Column AG reports the volume percentage of the first precipitate type in %.
- Column AH identifies the second precipitate type and assigns the same numerical labels as above.
- Column AI lists the size of the second precipitate type in nm.
- Column AJ reports the volume percentage of the second precipitate type in %.
- Column AK identifies the third precipitate type and assigns the corresponding numerical labels.
- Column AL lists the size of the third precipitate type in nm.
- Column AM reports the volume percentage of the third precipitate type in %.

• Columns AN through AW contain the mechanical properties reported for each material. Unless stated otherwise, all properties were obtained at room temperature.

- Column AN lists the ultimate tensile strength (UTS) in MPa.
- Column AO lists the ultimate compressive strength (UCS) in MPa.
- Column AP provides the tensile yield strength (TYS) in MPa.
- Column AQ provides the compressive yield strength (CYS) in MPa.
- Column AR lists the hardness in HV.
- Column AS gives the tensile ductility in %.
- Column AT gives the compressive ductility in %.
- Column AU lists the non-room-temperature test temperature, including both cryogenic and elevated temperatures.
- Column AV provides the non-room-temperature test-measured strength in MPa.
- Column AW provides the non-room-temperature test-measured ductility in %.

4. Descriptive Information of Extracted Features

4.1. Composition

Fig. 1 visualizes the extracted chemical compositions through a scatter plot arranged by element, with each point representing an individual composition entry. The plot highlights the frequent use of Fe, Ni, Co, Mn, and Cr as base elements, typically within 0.1–0.4 at. fraction; the

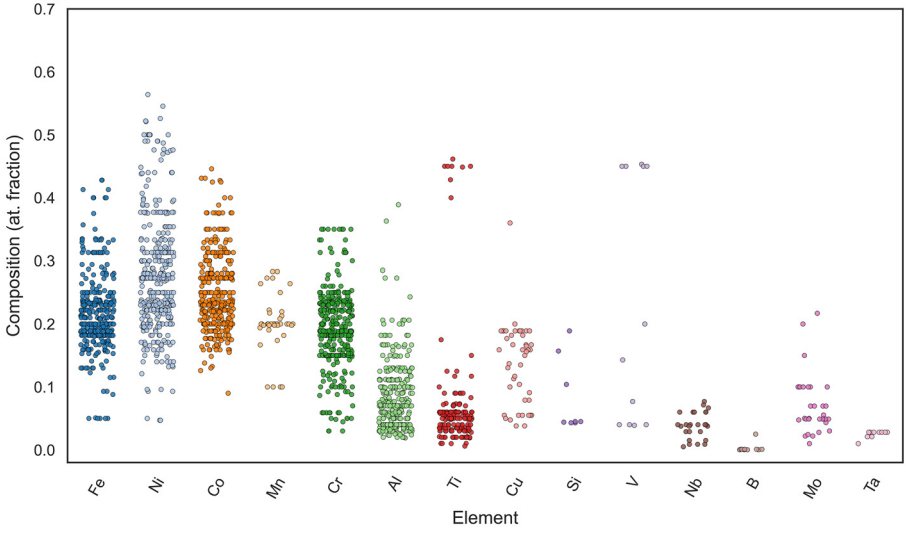

Fig. 1. Scatter plot of atomic-fraction compositions for 14 commonly reported chemical elements.

wider ranges of Al and Ti additions that influence ordering and strengthening; and the more targeted, lower-fraction use of elements such as Si, V, Nb, B, Mo, and Ta. This distribution underscores the compositional diversity and design strategies present in the reported literature, forming an essential foundation for subsequent analyses of processing pathways, microstructural characteristics, and properties.

4.2. Processing features

The processing history for each alloy was gathered, allowing us to consistently pull information on homogenization, rolling, recrystallization, and aging. Fig. 2 provides an overview of how often these processing steps appear in the dataset. As shown in Fig. 2a, homogenization is reported most frequently (67.5%), followed by aging (57.5%), rolling (55.6%), and recrystallization (48.4%).

Fig. 2 b and c show the range of thermal conditions for these heat treatments. Homogenization temperatures mostly fall between 1100 and 1250 ° C, while recrystallization temperatures cover a wider span, from around 600 ° C to more than 1200 ° C depending on whether the goal is recovery, grain growth, or full recrystallization. Aging treatments generally occur at lower

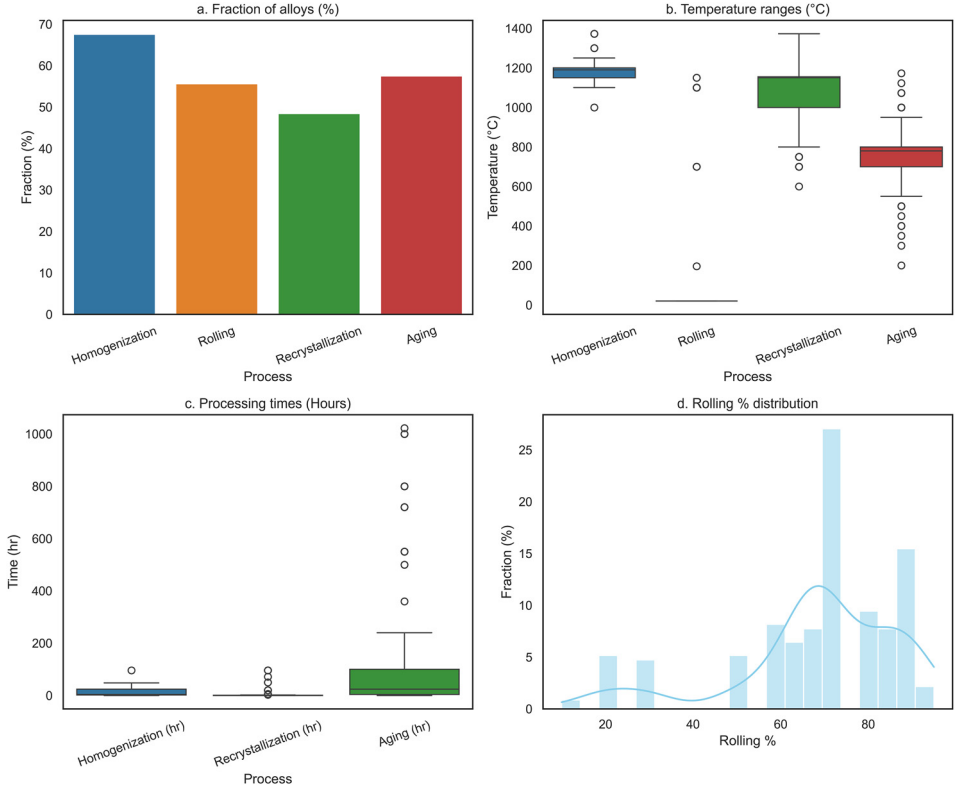

Fig. 2. Processing features. (a) Fraction of alloys subjected to homogenization, rolling, recrystallization, and aging. (b) Temperature ranges for each process. (c) Corresponding processing times. (d) Distribution of reported rolling reductions. Together, these plots summarize the processing routes used across the dataset.

temperatures (roughly 500–800 ◦ C), which fits their role in stabilizing phases such as B2, γ', or $D0_{19}$. Processing times vary just as widely — from brief exposures lasting under an hour to extended treatments exceeding 1000 hours — reflecting the different experimental approaches used to study phase stability and coarsening behavior.

Fig. 2 d shows the distribution of rolling reductions, which span a broad range but tend to cluster between 60% and 80%. This trend is consistent with typical MPEA studies, where moderate to heavy rolling is used to refine grains, introduce dislocations, or prepare the alloy for subsequent heat treatments.

Taken together, the processing data summarized in Fig. 2 capture the wide mix of thermal-mechanical routes reported in the literature and provide a useful basis for examining how these conditions influence microstructures and properties.

4.3. Microstructure features

In the dataset, although most matrices are single phase, dual-phase matrices are explicitly reported, reflecting the diversity of matrix states observed across different alloy systems.

Fig. 3 shows the overall distribution of precipitate phases gathered from the published studies included in this work. When there are no precipitates, the case is categorized as “none”. Only a small group of precipitate types appears regularly across the surveyed alloys. Among them, the $L1_2$ phase is the most prevalent, representing over 40% of all reported phases. Its prominence aligns with its well-known role as a key strengthening phase in many FCC-based MPEAs. The B2 phase follows as the next most frequently observed precipitate at roughly 20%, while FCC- and BCC-type precipitates each account for less than 10% of the total.

Several other phases — such as σ, η, $L2_1$, HCP, and μ — occur far less often, each contributing only a few percent. Rare phases including ε, γ'', γ', $D0_{19}$, and other ordered structures appear only intermittently in the literature, which may reflect narrow stability ranges or limited experimental investigation. Taken together, these trends show that most reported alloy systems still rely heavily on $L1_2$- and B2-based strengthening, while many other potential precipitate families remain comparatively unexplored.

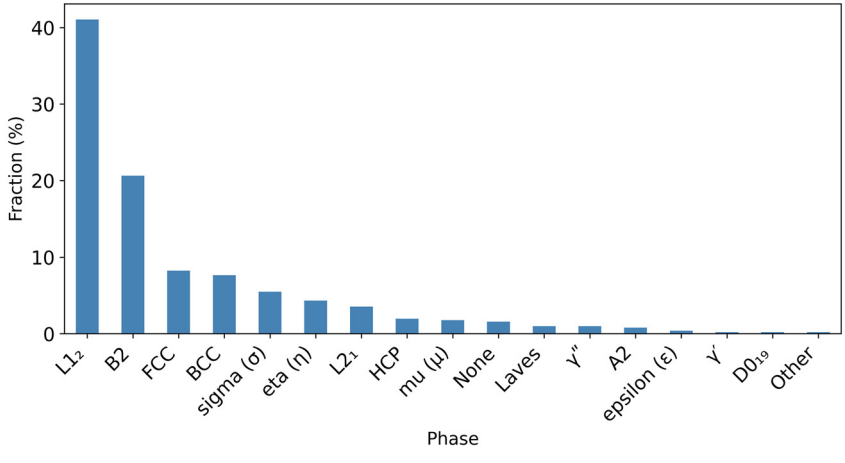

Fig. 3. Fractional distribution of reported precipitate phases. The category “none” means there are no precipitates in those combinations.

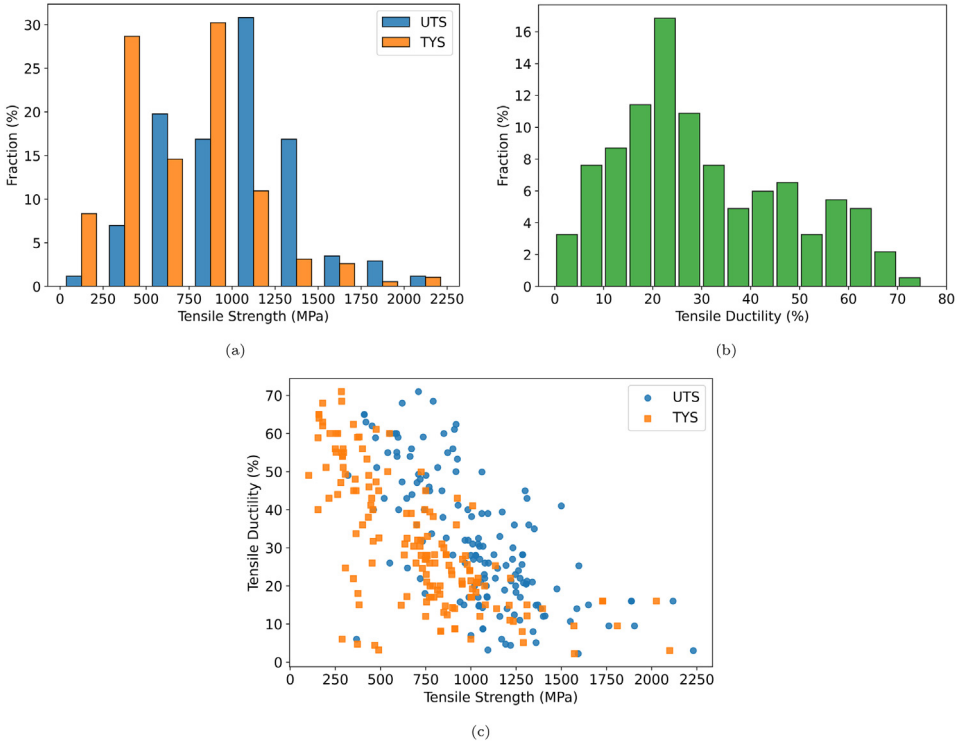

Fig. 4. (a) Fractional distribution of UTS and TYS values measured at room temperature. (b) Fractional distribution of tensile ductility values measured at room temperature. (c) Scatter plot of tensile ductility as a function of UTS and TYS at room temperature.

4.4. Property features

Figs. 4 and 5 illustrate the breadth of mechanical properties represented in the surveyed literature and the substantial variability characteristic of MPEAs. Fig. 4(a) summarizes the distributions of UTS and TYS gathered from these works. The values span a wide range, reflecting the breadth of alloy chemistries and processing paths explored in the literature. Most UTS measurements fall between 400 and 1300 MPa, with a noticeable concentration around 900–1100 MPa. TYS values show a similar overall pattern but appear at lower strengths, with clusters near 300–500 MPa and around 900 MPa. A small number of alloys reach well beyond 1500 MPa, and a few exceed 2000 MPa, demonstrating the level of strengthening that can be achieved in compositionally complex systems. For alloys where compressive data are available, Fig. 5a shows the corresponding distributions of UCS and CYS. These measurements likewise cover a broad range and generally follow the trends observed in tension, with CYS values consistently below UCS and a clear representation of high-strength alloys.

In addition to the strength data, the ductility distributions provide further insight into the range of deformation behaviors reported for precipitate-containing MPEAs. As illustrated in Fig. 4(b), tensile ductility values vary widely, from only a few percent to more than 70%, with the majority of measurements falling between 10% and 30%. The corresponding distribution of compressive ductility shown in Fig. 5b exhibits a higher proportion of alloys capable of sustaining large plastic strains, with many reported values above 40%. This behavior is consistent with the lower susceptibility to strain localization and fracture under compressive loading.

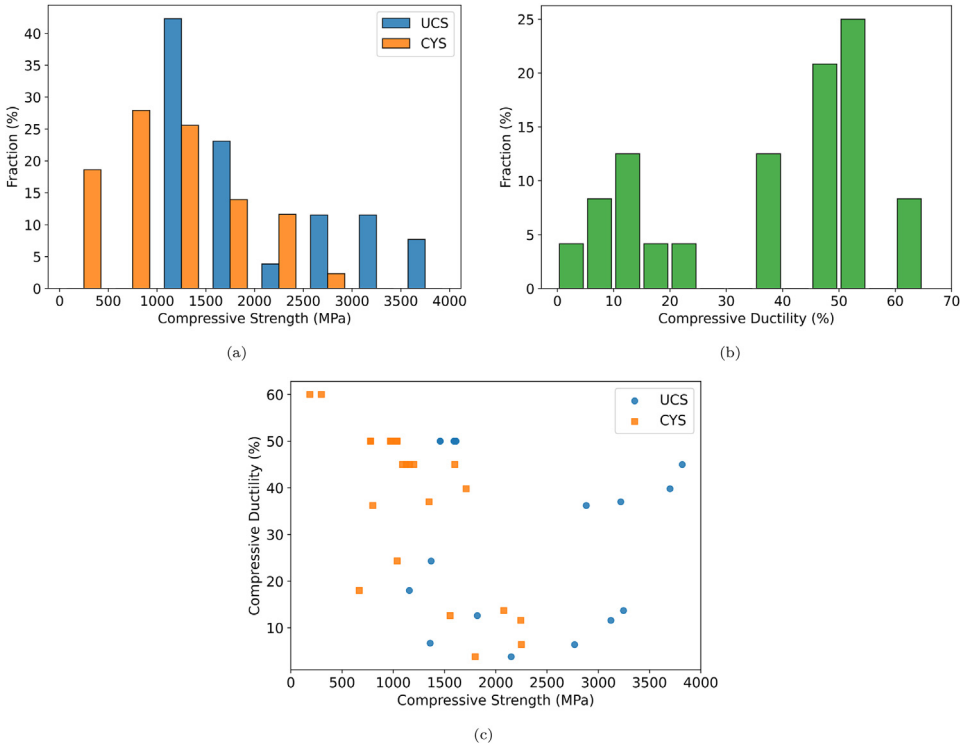

Fig. 5. (a) Fractional distribution of UCS and CYS values measured at room temperature. (b) Fractional distribution of compressive ductility values measured at room temperature. (c) Scatter plot of compressive ductility as a function of UCS and CYS at room temperature.

Compared with the tensile data, compressive data show a higher strength but comparable ductility. These emphasize the inherent tension-compression asymmetry in the mechanical response of MPEAs and highlight the need to consider loading mode when comparing strength and ductility across different studies.

Furthermore, the relationship between strength and ductility under tensile loading is shown in Fig. 4(c). As strength increases, ductility tends to fall, a trend commonly seen in structural alloys. Alloys with UTS below roughly 800 MPa often retain ductility above 40–50%, while many alloys with strengths above 1200 MPa show ductilities under 20%. Even so, several compositions manage to combine moderate or high strength with ductility above 30%. An analogous trend appears in compression, as illustrated in Fig. 5c, where compressive ductility decreases with increasing UCS and CYS, though a number of alloys still exhibit meaningful plasticity at intermediate strength levels.

The availability of mechanical property data in the literature is, however, uneven. Among the 418 entries in the full dataset, 218 alloys (52.15%) report at least one tensile property, while only 46 alloys (11.00%) include compressive strength or ductility. This imbalance reflects typical experimental practice in MPEA research, where tensile testing is far more common and compressive measurements are reported less frequently. As a result, the tensile dataset provides a more comprehensive basis for identifying trends, whereas the more limited compressive dataset still offers valuable insights but covers a smaller portion of the reported alloy space.

Limitations

The first limitation of our dataset is some missing features. For example, mechanical property measurement-related features such as strain rate and sample geometry, which can influence the interpretation of mechanical properties, are not included. As another example, volume fraction values for the matrix and precipitates were extracted directly as reported in the original publications, without distinguishing whether they were experimentally measured, estimated, or inferred, or specifying how they were obtained in the original articles. In addition, while the lattice types of precipitates are included, the chemical compositions of precipitates are not. Consequently, users are encouraged to consult the original sources for detailed information when interpreting or reusing our dataset.

Another limitation of our dataset is its relatively small number of entries. While the dataset includes 396 combinations of alloy composition and processing condition compiled from 100 peer-reviewed publications, it captures only a limited portion of the approximately 4360 studies on precipitate-containing MPEAs identified by searching the keyword — precipitate "multi-principal element alloys" — in Google Scholar in December 2025.

The relatively small size of the dataset is a result of the need for careful manual verification when high data fidelity is required. For example, complex microstructural descriptions, such as mixed phases, region-dependent reporting, or crystallographic notation that does not map cleanly onto predefined encoding schemes, often require domain expertise to interpret accurately. Similar considerations apply to the extraction of mechanical property data, where distinctions between experimentally measured values, inferred quantities, and theoretical discussions can be subtle and context dependent. While LLMs are effective at identifying and organizing relevant information from the literature, ensuring consistency and avoiding over-interpretation of qualitative descriptions benefits from expert review, which is usually labor intensive and costly. This trade-off highlights the complementary roles of automated extraction and expert validation in building trustworthy materials databases, highlighting to need to improve the robustness of LLM-based information extraction to enable the larger-scale experimental database for materials science.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Multi-Category Materials Information Extraction (Composition, Processing, Microstructure, Properties) (Zenodo) (Research Data).

CRediT Author Statement

Anshu Raj: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation; **Xin Wang:** Writing – review & editing, Investigation, Formal analysis, Data curation; **Matthew Luebbe:** Data curation; **Haiming Wen:** Funding acquisition, Supervision, Data curation; **Kun Lu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization; **Shuozhi Xu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Acknowledgments

A.R. and S.X. acknowledge the support of the U.S. National Science Foundation (DMREF-2522655). H.W. acknowledges the financial support by the U.S. Nuclear Regulatory Commission Faculty Development Program (award number NRC 31310018M0044). Financial support was provided by the University of Oklahoma Libraries' Open Access Fund.

Declaration of Competing Interest

The authors declare no competing interests.

References

[1] B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng. A 375–377 (2004) 213–218.
[2] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, S.Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (5) (2004) 299–303.
[3] B. Xu, H. Duan, X. Chen, J. Wang, Y. Ma, P. Jiang, F. Yuan, Y. Wang, Y. Ren, K. Du, et al., Harnessing instability for work hardening in multi-principal element alloys, Nat. Mater. 23 (6) (2024) 755–761.
[4] L. Liu, X. Liu, Q. Du, H. Wang, Y. Wu, S. Jiang, Z. Lu, Local chemical ordering and its impact on radiation damage behavior of multi-principal element alloys, J. Mater. Sci. Technol. 135 (2023) 13–25.
[5] E. Gienger, J. Rokisky, D. Yin, E.A. Pogue, B. Piloseno, A database of multi-principal element alloy phase-specific mechanical properties measured with nano-indentation, Data Brief 55 (2024) 110719.
[6] A. Zakutayev, N. Wunder, M. Schwarting, J.D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, An open experimental database for exploring inorganic materials, Sci. Data 5 (1) (2018) 180053.
[7] B. Lafuente, R.T. Downs, H. Yang, N. Stone, The power of databases: the RRUFF project, in: T. Armbruster, R.M. Danisi (Eds.), Highlights in Mineralogical Crystallography, De Gruyter, 2015, pp. 1–30.
[8] X. Wang, A. Raj, M. Luebbe, H. Wen, S. Xu, K. Lu, Reliable end-to-end material information extraction from the literature with source-tracked multi-stage large language models, 2025, 2510.05142, https://arxiv.org/abs/2510.05142
[9] E. Ma, J. Ding, Compositional fluctuation and local chemical ordering in multi-principal element alloys, J. Mater. Sci. Technol. 220 (2025) 233–244.
[10] Y. Sohail, C. Zhang, D. Xue, J. Zhang, D. Zhang, S. Gao, Y. Yang, X. Fan, H. Zhang, G. Liu, et al., Machine-learning design of ductile FeNiCoAlTa alloys with high strength, Nature 643 (2025) 119–124.
[11] X. Wang, A. Raj, M. Luebbe, H. Wen, S. Xu, K. Lu, Multi-category materials information extraction (composition, processing, microstructure, properties), 2025, 10.5281/zenodo.17612940