scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High Entropy Alloys Database generated with Large Language Model

Vladimir Chizhevskiy[1,2], Gordana Marković[3,4], Salah-eddine Benrazzouq[4], Uroš Cvelbar[1,2] ✉, Alexandre Nominé[1,4] ✉ & Janez Zavašnik[1,2,5] ✉

High entropy alloys (HEAs) represent a promising area in materials science, but systematic analysis of the extensive literature remains a challenge. In this study, we used Natural Language Processing (NLP) techniques to analyze 4,625 scientific articles from a restricted corpus representing publisher-accessible literature, successfully identifying and characterizing 12,427 of different high entropy alloys. Through prompt engineering and experiments with Large Language Models (LLMs), including mamba-transformer hybrid architectures, we developed a structured database that captures important parameters such as alloy compositions, phase numbers and crystallographic structures. In our analysis, we distinguish between theoretical and experimental studies, considering specific methodological details for each category. For theoretical work, we have systematically documented modeling approaches and key computational parameters, while experimental studies are cataloged with their synthesis methods and critical processing conditions. This database represents a large-scale, automated extraction of HEA research data. The accuracy of the data ranges from 78.7% for HEA phase identification to 94.3% for HEA composition.

Background & Summary

The green and digital transition will generate an unprecedented demand for materials[1,2]. Serious concerns have emerged regarding our ability to meet these material demands using existing resources, as several critical elements—including Platinum group metals, Rare Earth elements, Gallium, and Germanium—already face significant supply risks[3]. One promising strategy involves developing novel substitute materials, particularly in the form of advanced alloys. The potential solution space is extraordinarily vast; the number of possible alloy combinations is astronomically large[4], making traditional trial-and-error discovery approaches highly inefficient. Artificial Intelligence (AI) is a promising avenue to accelerate this materials discovery process. However, the effectiveness of AI-driven discovery depends critically on access to extensive and reliable materials databases. Historically, materials databases have been developed through either manual curation of experimental results[5] or computational modeling approaches[6]. Manually curated databases contain highly reliable information extracted from peer-reviewed scientific literature, representing experimentally validated results. However, these databases suffer from limited coverage and require time-consuming development and maintenance. Conversely, computational databases like the Materials Project (https://www.materialsproject.org/) offer much broader coverage but face questions of real-world applicability, as theoretical predictions frequently diverge from experimental outcomes. A significant advancement came with tools like ChemDataExtractor[7], which can analyze tens to hundreds of thousands of scientific papers to automatically extract valuable physical properties such as yield stress[8], Curie and Neel Temperatures[9], refractive indexes[10], bandgap[11], thermoelectric figure of merit[12], etc. To further enhance the power of AI-based materials discovery methods, additional descriptors and more comprehensive property extraction are needed.

In this paper, we focus on High Entropy Alloys (HEAs) as a case study, given their promise in various applications relevant to the green and digital transition, including exceptional mechanical properties[13,14] and catalytic

[1]Jožef Stefan Institute, Department of Gaseous Electronics, Jamova cesta 39, 1000, Ljubljana, Slovenia. [2]Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000, Ljubljana, Slovenia. [3]Institute for Technology of Nuclear and Other Mineral Raw Materials, Bulevar Franš d'Eperea 86, Belgrade, Serbia. [4]Université de Lorraine, CNRS, IJL, 2 allée André Guinier, Nancy, France. [5]Max Planck Institute for Sustainable Materials, Max-Planck-Str. 1, 40237, Düsseldorf, Germany. ✉e-mail: uros.cvelbar@ijs.si; alexandre.nomine@univ-lorraine.fr; janez.zavasnik@ijs.si

Publisher	Number of documents
Elsevier	4,248
Springer Nature	72
The Royal Society of Chemistry	305

Table 1. Distribution of Scientific Papers in the Dataset per Publisher.

performance[15–17], and their characteristically broad compositional space. Using Large Language Models, we have extracted crucial data necessary for AI-based exploration of HEAs, such as phase identification, crystallography, and chemical composition. To accelerate future experimental development, we have also included the extraction of synthesis methods and theoretical modeling approaches described in the scientific literature, establishing a more comprehensive foundation for accelerated materials discovery. Both components, (1) the HEA dataset, and (2) the LLM-driven extraction methodology, constitute the main scientific contribution, with the dataset being the primary scientific output and the methodology being necessary to generate it.

Methods

Data Acquisition. According to Web of Science (WoS), 22,745 English-language documents on high-entropy alloys have been published since 2004; approximately 80% of these are not accessible through the APIs used in this study. From the publicly accessible restricted corpus, we collected 4,625 scientific papers, representing about 20% of the total HEA publications. The distribution of these papers across scientific publishers is shown in Table 1.

This restricted corpus was acquired through full-text articles retrieval via various publisher APIs (Application Programming Interface) such as Elsevier's TextMining API (https://dev.elsevier.com), representing a publisher-accessible subset of the broader HEA literature. A significant proportion of the research papers were obtained in the XML format (eXtensible Markup Language), which provides structured semantic markup suitable for LLM-based processing pipelines. For publications sourced from alternative publishers (Royal Society Of Chemistry (https://developer.rsc.org) and Springer (https://www.springernature.com)), which were predominantly available in PDF (Portable Document Format), we implemented a machine learning model trained to transform complex scientific PDF layouts into markdown syntax. Only factual numerical data and short descriptors were extracted; no copyrighted figures, tables, or narrative text were redistributed. Document retrieval requires special authentication available to the research institutions and unique identifiers for precise targeting, using keyword queries 'high entropy alloy' or 'HEA' applied to titles, abstracts, and full texts. These identifiers take the form of either:

- Digital Object Identifiers (DOIs), or
- Platform-specific unique identifiers.

For publications from the Royal Society of Chemistry (RSC), the temporal parameter (publication year) served as an additional mandatory retrieval criterion.

To translate some of the papers which were available only in PDF format to markdown, we had to employ one more language modeled trained just for that: llama-parse from llama-index converts structured PDFs with its figures, tables, formulas, etc. into the markdown.

Dataset generation. In order to construct a comprehensive database of high-entropy alloys from publicly available scientific literature, we developed a systematic extraction pipeline. This methodology involved sequentially processing individual scientific articles in various text formats through a large language model. Each paper was loaded into the context window of the model, followed by a series of structured queries. Importantly, the model retained its previous responses during the questioning sequence, enabling consistent information extraction across the entire document corpus and facilitating the construction of a coherent, structured dataset.

Transformer-based architectures currently dominate the field of large language models (LLMs), with prominent examples including most open-source models, such as Llama, and many proprietary models, such as those from the GPT-family. Recent advances in hybrid architectures, which combine Transformers with Mamba sequence modeling and Mixture-of-Experts (MoE) approaches, have been applied to handling extensive context windows. These models maintain competitive inference times while processing substantial inputs such as complete books or comprehensive scientific journals. In our analysis, we utilized Jamba 1-5 Large, which employs 94 B active parameters from a total parameter count of 398 B. This substantial difference between active and total parameters is inherent to MoE-based[18] architectures, where only a subset of experts is activated for each input token, allowing for efficient computation while maintaining model capacity.

The concept of utilizing Mamba for sequence modeling in a linear-time fashion, which is efficient for managing large volumes of data without compromising on response time, has been discussed extensively by Waleffe *et al.*[19]. Moreover, the specific implementation details and empirical results for these hybrid models have been further elaborated by Lieber *et al.*[20], who demonstrated the utility of Mamba-based architectures in large-scale language modeling. Additional insights into the architecture and performance of hybrid models can be gleaned from the studies conducted by the Jamba team and Lenz *et al.*[20,21], which discuss the structural nuances and operational efficiencies of these recent model designs. This approach was selected for the proposed data generation pipeline because it is suitable for processing entire scientific articles within a single context window, enabling consistent information extraction across the document.

Data	Short Summary	Data Type
DOI	Source article DOI	String
Prompt 1 answer	Identify alloys	String
Prompt 2 answer	Determine the crystallographic phases	String
Prompt 3 answer	Check type of material prediction	String
Prompt 4 answer	Check presence of partially ordered structure	String
Prompt 5 answer	Create comprehensive JSON structure with the results	String

Table 2. Schema for the table of HEA-related scientific papers.

Prompt Engineering and Context Management. The design and implementation of prompting mechanisms play a crucial role in maximizing the efficacy of Large Language Model (LLM)-based information extraction pipelines.

Well-constructed prompts act as control parameters that enable structured extraction, standardised formatting, and retrieval of relevant scientific information from research literature. An example of a prompt for the data extraction:

"For each high entropy alloy [$X_1X_2...X_n$] identified in the paper, determine the crystallographic phases present. Describe the nature of each phase: amorphous, solid solution, intermetallic compound. Include chemical formula, crystallographic state, structure, and space group if available. Note any precipitates that may not be visible in X-ray diffraction but are evident in micrography."

In our methodology, we implemented an iterative chain-of-thought prompting approach, where in each subsequent interaction with the LLM incorporated the cumulative context from all previous exchanges:

$$P_n = f(Q_n, \{(Q_1, R_1), (Q_2, R_2), \ldots, (Q_{n-1}, R_{n-1})\}) \quad (1)$$

where:

- P_n represents the n-th prompt,
- Q_n is the current query,
- (Q_i, R_i) represents the question-response pair from the i-th interaction,
- f is the context aggregation function that combines the current query with historical interactions.

This recursive prompting strategy enables the model to maintain contextual continuity while building upon prior analyses, enabling accumulation of contextual information across interactions. The preservation and utilization of conversation history allows interpretation of complex materials science concepts and relationships, as each prompt benefits from the contextual enrichment of preceding interactions.

The implementation followed a structured pipeline:

- **Initial Context Formation**: Base prompts were designed to establish fundamental material properties and relationships,
- **Iterative Refinement**: Each subsequent prompt incorporated previous responses through a context window,
- **Response Validation**: Automated verification of response format and content consistency. This can be achieved by specifying parameters responsible for providing a structured output: in the final prompt, the model receives the detailed structure presented as nested Python classes or JSON dictionary schema as one of the parameters of the request. This way, the model is obliged to return the data in a specified semi-structured format,
- **Context Management**: Dynamic adjustment of context window to maintain relevant information while avoiding token limits.

To construct our final database[22], we use 5 prompts, each designed to identify a different key component of the analysis. While the released HEA dataset[22] is the primary outcome of this work, the LLM-driven extraction methodology enables dataset creation at a larger scale than manual curation.

Data Records

In the course of sequential querying of scientific articles followed by the parsing of nested JSON-like strings generated by the language model, two distinct tables have been constructed.

The first table encompasses a collection of High Entropy Alloys (HEA)-related scientific papers, encapsulating all responses to the queries outlined in the section Summary of Data Records in Table 2. The second table compiles a detailed inventory of alloys as described within these articles, systematically organized into a tabular format (Summary of Data Records in Table 3). The second table is our targeted final data: a database of high entropy alloys (HEAs)[22], extracted from 4,625 peer-reviewed scientific publications. Each database row contains the DOI of the source publication, enabling unambiguous traceability of every extracted alloy record.

Both datasets are available for download on Mendeley Data[22].

Data	Description	Data Type
Alloy	Formula of the alloy	String
Article name	Source article title	String
Number of phases	N of crystallographic phases	Integer
Experimental or theoretical	Type of details provided	String
Experimental details	Details of experiment	String
Theoretical details	Theoretical details	String
Special conditions	Conditions applied in experiment	String
Crystallographic structure	Predicted structure of HEA	String
DOI	Source article DOI	String
Journal	Published journal	String

Table 3. Schema for the resulting database of high-entropy alloys.

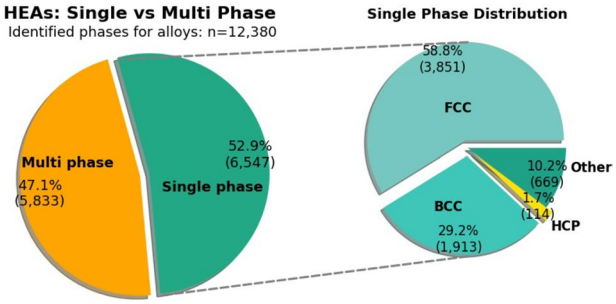

Fig. 1 Single phase HEA distribution in the resulting database.

Each dataset[22] row contains the DOI of the source publication, enabling unambiguous traceability of every extracted alloy record.

Data Overview. *Crystal Structure Preference.* Among single-phase alloys, FCC (face-centered cubic) structures dominate at 58.8% (3,851 alloys), followed by BCC (body-centered cubic) at 29.2% (1,913 alloys), and HCP (hexagonal close-packe d) representing only 1.7% (114 alloys) (Fig. 1). The remaining 10.2% (669 alloys) fall within other crystallographic structures. This distribution reflects the strong preference for FCC and BCC structures in HEA design strategies, likely due to their favorable combination of strength and ductility.

A detailed analysis of the fabrication techniques compiled in our database reveals clear trends in the synthesis approaches used for high-entropy alloys (HEAs) (Fig. 2). We ran a script to extract information on fabrication techniques for all of the HEA alloys (n=12,427). The traditional metallurgical techniques in HEA development are the most common, and melting methods comprise 69% (n=8,580) of all documented fabrication routes. Among them, arc melting accounts for 60% (n=5,175) of all melting processes, which is primarily due to its accessibility, cost-efficiency, and capacity to achieve uniform elemental mixing under a controlled atmosphere.

Mechanical processing is the second most common approach, used in 24.8% (n = 3,086), with ball milling as the most common method by accounting for 77% (n = 2,372) of the cases. This shows the importance of powder metallurgy routes in HEA research, particularly in terms of microstructure refinement and the introduction of elements with largely differing melting points. Other specialized processing methods account for 25.2% (n=3,126) of the entire database, with general processing methods comprising 78% (n = 2,451) of this category. Falling into this category are extensive applications of powder metallurgy and overall references to additive manufacturing, prevailing in cases where exact technique details were not specified within the original work. This large percentage also reflects the role of secondary manufacturing techniques, thermal processing and post-processing treatments in the development of HEAs. It can be observed that the uptake levels of novel technologies are considerable yet remain dynamic. Laser-based techniques account for 10.6% of the total fabrication processes. The comparatively small presence of sputtering techniques (3.4%) and thermal spray techniques (9.5%) indicates that these techniques are primarily used for specialized purposes, i.e., coating technologies and surface engineering.

Elemental Composition Trends. The periodic table heat map (Fig. 3) demonstrates clear preferences in element selection for HEA design. Transition metals show notably high utilization, with Cr (68.9%), Ni (69.8%), Fe (65.5%), and Co (62.5%) as the most frequently incorporated elements, with significant representation of Al (36.6%). These element selections demonstrate the focus of researchers on compositions around the Cantor alloy (CoNiCrFeMn) and, subordinately, around refractory metals.

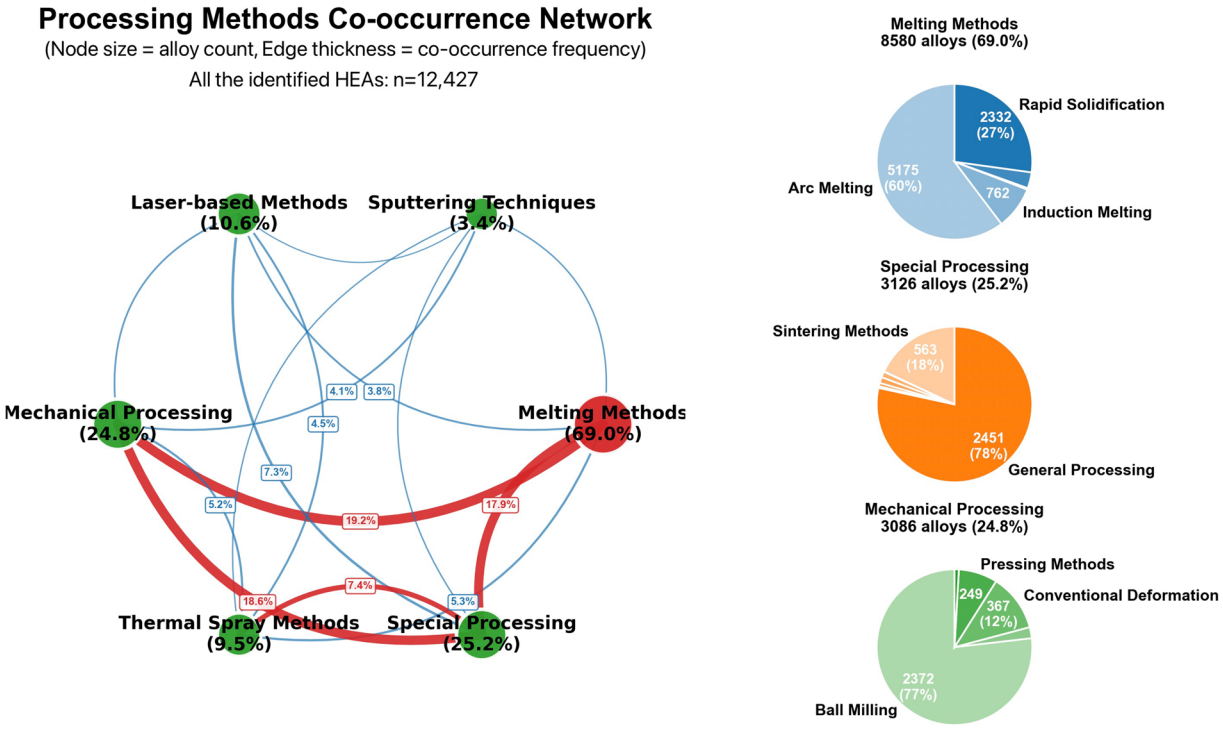

Fig. 2 Distribution of fabrication methods in the HEA database with hierarchical breakdown. The network chart shows the six primary categories (Node size = alloy count, Edge thickness = co-occurrence frequency), while smaller charts detail the subcategory distributions within 3 of the most popular processing route.

Technical Validation

LLM-generated results and the results parsed by the script. The discrepancy between evaluation metrics results from the two-stage data processing pipeline. To reflect the two-stage structure of the pipeline, two complementary evaluation metrics were used. The extraction score measures the model's ability to correctly identify and interpret materials-related information from the source text, while the parsing score assesses whether the generated output conforms to the predefined JSON schema required for downstream tabularization. This separation allows performance losses to be attributed either to extraction errors or to inconsistencies introduced during post-processing. Following the extraction of JSON-structured data from scientific literature by the large language model (LLM), a post-processing script transformed these semi-structured data objects into a tabular format suitable for the resulting relational database. Each row in the final table represents a distinct alloy system described in the original publications, with associated microstructural and compositional information. The parsing of these extracted JSON dictionaries presented significant challenges due to inconsistencies in the LLM's output format, despite the provision of exemplar templates during the prompt engineering phase. A representative example of such inconsistency is found within compositionally graded alloy systems. When encountering studies on high-entropy alloys such as AlCoCrFeNi(TiN)x, where x represents variable stoichiometric ratios, the LLM frequently generated a single database entry with "AlCoCrFeNi(TiN)x" in the alloy field and aggregated all phase information from the various compositions into this single entry. The correct representation would have been separate database entries for each discrete value of x, each with its corresponding phase information. These structural inconsistencies in the extracted data necessitated the implementation of two distinct evaluation metrics: (1) an extraction quality score that quantifies the LLM's ability to correctly identify and extract relevant information from the scientific text, and (2) a parsing quality score that measures how effectively the generated data conforms to a standardized database schema and whether it is compatible with further data transformations. While the extraction score evaluates the accuracy of information retrieval from a materials science perspective, the parsing score specifically assesses data consistency and compatibility with the structured database requirements. This dual scoring system highlights the performance differences between content extraction accuracy and data format adherence in automated materials data extraction systems.

Validation process. The nature of the data set is distinctly unique, making conventional cross-validation infeasible. Consequently, substantial engagement from subject matter specialists is essential. To assess the plausibility and scientific validity of the Large Language models employed in this research, domain experts in materials science and metallurgy manually evaluated the derived predictions. This evaluation helps enhance the model's outputs and in pinpointing potential areas for improvement in the retrieval and generation processes. The experts initially established a reference library containing accurate responses for 50 randomly selected publications on HEA systems. The evaluation examined both the findings generated by the LLM and the post-processing script that converts the semistructured JSON data into a tabular format for the relational database. The

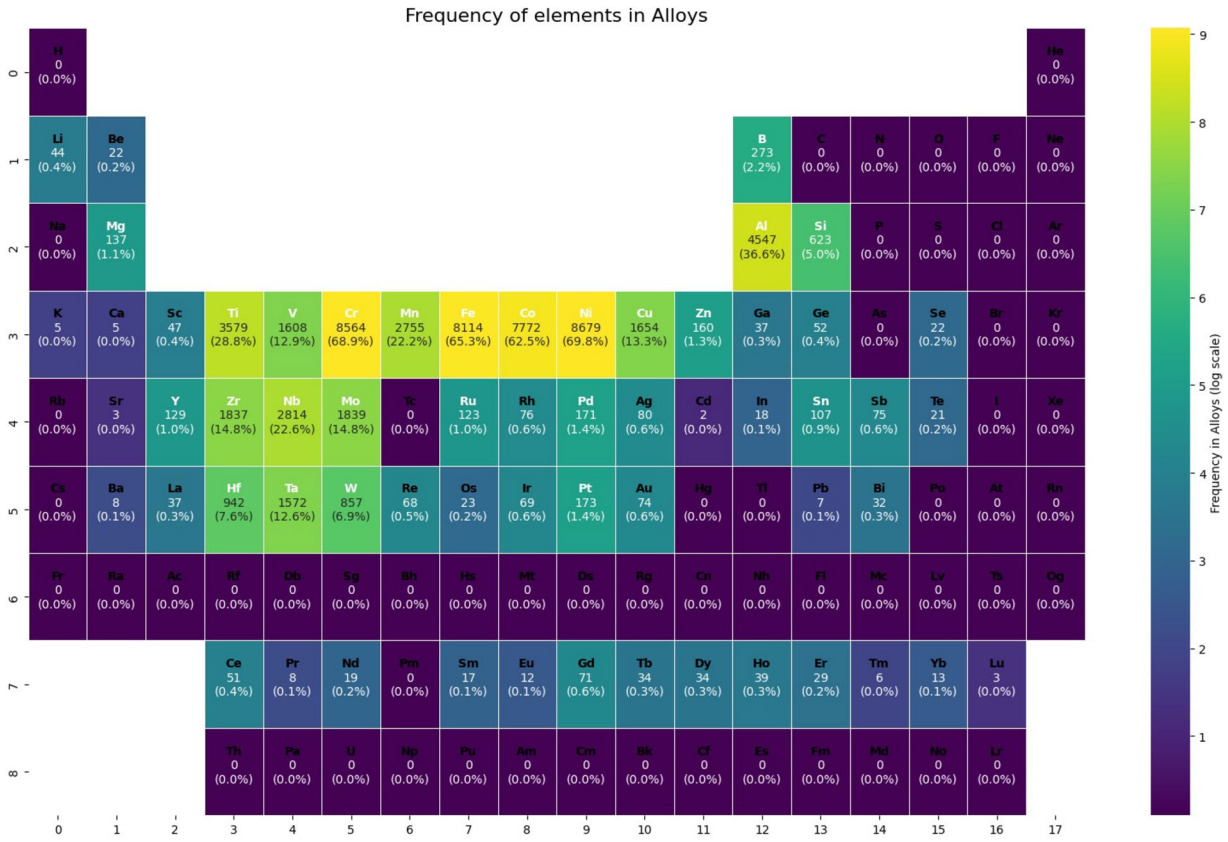

Fig. 3 Frequency of Elements in Alloys presented as Mendeleev's Periodic Table.

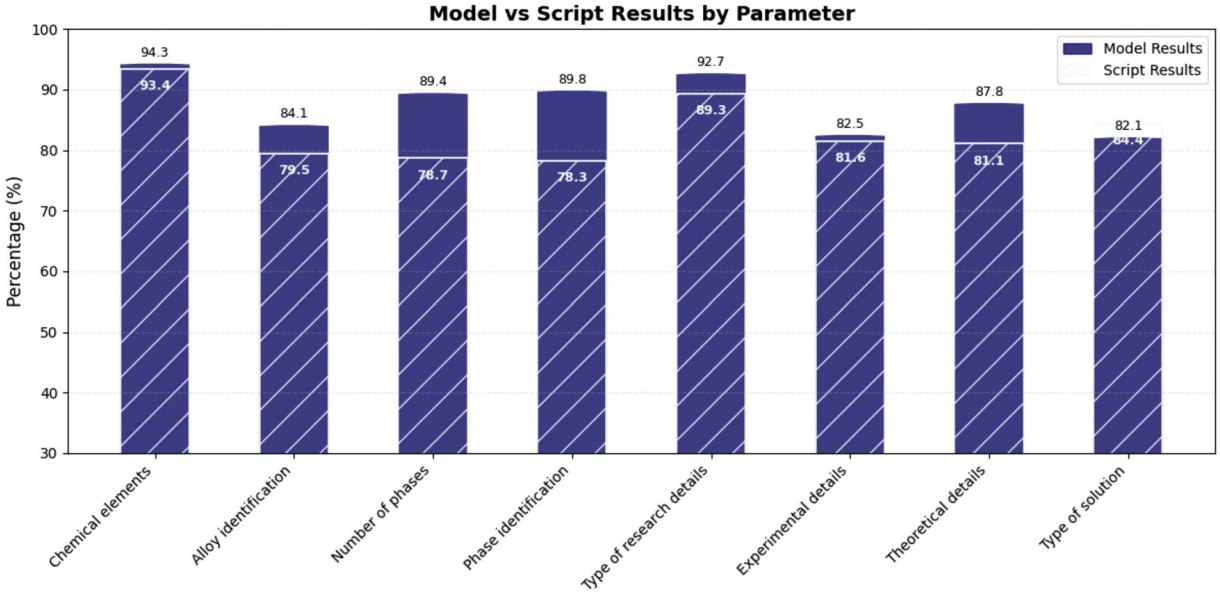

Fig. 4 Model vs Script results by parameter.

expert-generated responses were subsequently compared with the model outputs and the script-analyzed results. The evaluation concentrated on essential information extraction tasks, encompassing the accurate identification of components within chemical formulas, the precision of the formula, the quantity and identity of the reported phases, the synthesis type (experimental, theoretical or both), the extracted synthesis parameters, and the categorization of solid solution types. Binary scoring (1 for accurate, 0 for erroneous) was used where relevant, such as to evaluate the number of phases. A graded scoring approach was used for more nuanced output,

assigning 0 for incorrect, 0.5 for mostly accurate, and 1 for correct responses. The assessment indicated that the script shows somewhat lower performance than the model in most parameters, which is expected since parsing creates additional opportunities for errors. High-performance parameters prominently encompass the chemical formula, study type specifics, and experimental details, indicating that these elements represent well-organized data suitable for conversion into tabular format.

Parameters demonstrating significant performance deterioration (Fig. 4), including Number of Phases, Phase Identification, and Theoretical Information, signify difficulties in the conversion from JSON format. The script shows enhanced performance for "Type of solution" (84.4% compared to 82.1%), due to programmatic adjustments that enabled the script to leverage additional contextual information from the JSON dictionary to correct previously misclassified solution types. The assessment revealed that while the script typically incurs a loss of accuracy during parsing (average decrease of about 5-10%), the conversion to tabular format is crucial for database operation. The most notable decline in performance occured in "Phase identification" parsing, decreasing from 89.8% to 78.3%, likely due to uneven formatting of concentration values in the source JSON that required interpretive processing.

Because the main purpose of the manuscript is Data Descriptor, we did not conduct downstream predictive modeling, benchmarking, or task-specific performance evaluations. Instead, the dataset is provided as a resource to facilitate such analyses by future users. Accordingly, this work focuses solely on data acquisition, extraction methodology, and validation of data consistency.

Data availability

The database of high-entropy alloys and all supplementary information can be openly accessed via Mendeley Data[22] under https://doi.org/10.17632/j75v9bbbjz.

Code availability

The source code of this project used to generate the dataset is openly available on Project's GitHub: https://github.com/Vladimirchizh/hea_database.

Received: 28 August 2025; Accepted: 17 February 2026;
Published online: 05 March 2026

References

1. King, A. H. Our elemental footprint. *Nat. Mater.* **18**, 408–409 (2019).
2. Vidal, O., Goffé, B. & Arndt, N. Metals for a low-carbon society. *Nat. Geosci.* **6**, 894–896 (2013).
3. Lèbre, E. *et al*. The social and environmental complexities of extracting energy transition metals. *Nat. Commun.* **11**, 4823 (2020).
4. Cantor, B. Multicomponent high-entropy Cantor alloys. *Prog. Mater. Sci.* **120**, 100754 (2021).
5. Machaka, R., Motsi, G. T., Raganya, L. M., Radingoana, P. M. & Chikosha, S. Machine learning-based prediction of phases in high-entropy alloys: A data article. *Data Brief* **38**, 107346 (2021).
6. Chen, W. *et al*. A map of single-phase high-entropy alloys. *Nat. Commun.* **14**, 2856 (2023).
7. Swain, M. C. & Cole, J. M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
8. Kumar, P., Kabra, S. & Cole, J. M. Auto-generating databases of yield strength and grain size using ChemDataExtractor. *Sci. Data* **9**, 1–11 (2022).
9. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5**, 180111 (2018).
10. Zhao, J. & Cole, J. M. A database of refractive indices and dielectric constants auto-generated using ChemDataExtractor. *Sci. Data* **9**, 1–11 (2022).
11. Dong, Q. & Cole, J. M. Auto-generated database of semiconductor band gaps using ChemDataExtractor. *Sci. Data* **9**, 1–11 (2022).
12. Sierepeklis, O. & Cole, J. M. A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor. *Sci. Data* **9**, 1–12 (2022).
13. Lei, Z. *et al*. Enhanced strength and ductility in a high-entropy alloy via ordered oxygen complexes. *Nature* **563**, 546 (2018).
14. Fu, Z. *et al*. A high-entropy alloy with hierarchical nanoprecipitates and ultrahigh strength. *Sci. Adv.* **4**, eaat8712 (2018).
15. Batchelor, T. A. A. *et al*. High-entropy alloys as a discovery platform for electrocatalysis. *Joule* **3**, 834–845 (2019).
16. Broge, N. L. N., Bondesgaard, M., Søndergaard-Pedersen, F., Roelsgaard, M. & Iversen, B. B. Autocatalytic formation of high-entropy alloy nanoparticles. *Angew. Chem. Int. Ed Engl.* **59**, 21920–21924 (2020).
17. Li, H. *et al*. Fast site-to-site electron transfer of high-entropy alloy nanocatalyst driving redox electrocatalysis. *Nat. Commun.* **11**, 5437 (2020).
18. Shazeer, N. *et al*. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. Preprint at https://arxiv.org/abs/1701.06538 (2017).
19. Waleffe, R. *et al*. An empirical study of mamba-based language models. Preprint at https://arxiv.org/abs/2406.07887 (2024).
20. Lieber, O. *et al*. Jamba: A hybrid transformer-mamba language model. Preprint at https://arxiv.org/abs/2403.19887 (2024).
21. Team, J. *et al*. Jamba-1.5: Hybrid transformer-mamba models at scale. Preprint at https://arxiv.org/abs/2408.12570 (2024).
22. Zavašnik, J. *et al*. HEA_database Mendeley Data, https://doi.org/10.17632/j75v9bbbjz (2025).

Acknowledgements

The authors acknowledge the financial support from the European Innovation Council Pathfinder project under ThermoDust grant agreement No. 101046835. JZ and UC acknowledge the support by Slovenian Research Agency (ARRS) program P1-0417 and project J2-4440. Project is partially funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, reference number 451-03-136/2025-03/200023, and the Erasmus+ Traineeship Programme.

Author contributions

All authors contributed equally to the conception and design of the work. VC, GM, and SB carried out the acquisition of the data. All authors contributed equally to the analysis and interpretation of data. All authors have contributed to drafting, writing, and revising the work.

Competing interests

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

Additional information

Correspondence and requests for materials should be addressed to U.C., A.N. or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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