MATSEEK: AN AUTOMATED KNOWLEDGE-DRIVEN FRAMEWORK FOR MATERIALS RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

The discovery of advanced alloy materials increasingly depends on reliable and interpretable knowledge extracted from the scientific literature to guide data-driven composition–property optimization. While large language models (LLMs) have enabled automated database construction, existing approaches typically separate data extraction from relational scientific knowledge mining, limiting interpretability and physical grounding in materials design. Here we present **MatSeek**, an LLM-based framework that unifies structured alloy data and literature-derived scientific knowledge. MatSeek combines an automated pipeline for building high-quality alloy databases with a knowledge extraction module capturing empirical trends, mechanistic insights, and composition design principles. This knowledge can effectively accelerate machine-learning–driven alloy discovery by constraining exploration of composition space, while providing mechanistic explanations for model predictions. Applying MatSeek to 10,240 high-entropy alloy publications, we construct a database of 27,438 records and demonstrate efficient, interpretable identification of promising alloy compositions. MatSeek establishes a unified, literature-grounded paradigm for knowledge-driven materials discovery.

1 INTRODUCTION

Alloys form the foundation of modern materials engineering by providing tunable mechanical, electrical, and durability properties essential to next-generation aircraft and automobiles, energy infrastructure, electronics and other diverse technological applications (Pollock, 2016; Han et al., 2024). Recent artificial intelligence (AI)-driven approaches (Cheng et al., 2026) reshape the landscape of designing new alloy materials (Tshitoyan et al., 2019; Rickman et al., 2019; Pei et al., 2024; Gruver et al., 2024; Yang et al., 2025; Deng et al., 2025), advancing the traditional trial-and-error experimentation. In particular, large-scale and well-characterized databases improve the accuracy and reliability of materials inference based on this AI-driven paradigm. However, existing databases are predominantly constructed through manual literature curation (Rickman et al., 2019; Gorsse et al., 2018; Borg et al., 2020) or rule-based (Swain & Cole, 2016; Yan et al., 2022; Pei et al., 2023; Li et al., 2023; Wang et al., 2022) approaches that are labor-intensive, difficult to generalize across diverse alloy systems, and brittle to variations in publication styles. Automated materials information extraction has increasingly become a necessity.

Recent advances in large language models (LLMs) have provided an opportunity for the automated extraction of scientific knowledge from the materials literature (Pfeiffer et al., 2022; Dagdelen et al., 2024; Zhang et al., 2023; Foppiano et al., 2024; Choi & Lee, 2024; Polak & Morgan, 2024). Most of the existing studies primarily focus on extracting data-centric information, such as alloy compositions and properties, to construct structured databases. These LLM-derived databases provide a scalable and low-cost foundation for training machine-learning (ML) models. Beyond data extraction, growing efforts have explored the use of LLMs to capture relational and conceptual knowledge, including structure–property relationships and underlying physical principles (Bai et al., 2025; Yang et al., 2022; Mostafa et al., 2024; Pruyn et al., 2025; Durmaz et al., 2024), enabling powerful AI assistants for materials discovery. In fact, both structured data information and relational scientific knowledge are inherently interconnected since they are derived from the same textual sources. However, existing approaches typically extract and utilize them in isolation, without explicitly modeling their mutual dependencies. This separation obscures the links between empirical observations and the underlying scientific reasoning, leaving materials discovery workflows that are hard to interpret,

rigorously validate, and trust. To address this challenge, this work proposes a unified framework that integrates structured data extraction with relational knowledge mining from materials literature, enabling physics-informed guidance of ML–driven optimization while providing mechanistic explanations for the resulting design recommendations.

Here, we propose MatSeek, an LLM-based materials design framework that unifies structured data and relational scientific knowledge within a single workflow. Specifically, MatSeek is built around a data extraction pipeline, termed Auto-Data Pipeline, and a knowledge extraction module, which together support an ML-driven materials discovery process. The Auto-Data Pipeline automatically extracts data information from scientific articles, such as compositions, processing procedures, and reported properties, to construct a structured materials database. In parallel, the Knowledge extraction module identifies scientific knowledge from the literature, including empirical trends, mechanistic insights, and composition design criteria. Then, the ML-driven materials discovery module integrates these two outputs: ML predictors are trained on the structured data generated by the Auto-Data Pipeline, while the search for inverse design is explicitly guided by constraints derived from the extracted scientific knowledge. Finally, the literature-derived knowledge is used to provide mechanistic interpretations of the candidate materials. In this way, MatSeek enables efficient and physically meaningful exploration of composition space, bridging data-centric learning and knowledge-driven reasoning.

To demonstrate the effectiveness of MatSeek, we first evaluate the performance of its Auto-Data Pipeline and demonstrate its ability to reliably construct alloy databases. Applying this pipeline to 10,240 publications on high-entropy alloys (HEAs), we assemble a large-scale database comprising 27,438 data records. We train an ML model based on this database for property prediction and incorporate semantic knowledge to guide efficient exploration of composition space for targeted alloy design. These case studies illustrate how MatSeek systematically integrates structured data with relational scientific knowledge extracted from the literature, enabling efficient and interpretable materials discovery within a unified, literature-grounded framework.

2 Results

2.1 Overview of the MatSeek framework

Figure 1 presents an overview of MatSeek, from auto-data pipeline and knowledge mining, knowledge-guided material discovery, to mechanistic interpretation with knowledge graph. The system first processes scientific publications to extract alloy compositions, processing routes, testing conditions, and property information, which are consolidated into a structured materials database. This database provides the foundation for training ML predictors of alloy properties. In parallel, MatSeek extracts high-level scientific knowledge from the literature, including relational knowledge, material trends, and design principles. The trained predictors are then coupled with optimization algorithms, where the search over the vast composition space is explicitly guided and constrained by the extracted scientific knowledge. To validate the reliability of the ML model predictions, the resulting candidate compositions are assessed using thermodynamic simulation software, such as Thermo-Calc (Andersson et al., 2002). Then, the compositions are interpreted using a LLM operating over the constructed knowledge graph. Overall, MatSeek unifies structured data and scientific knowledge within a single workflow, enabling efficient, interpretable, and knowledge-grounded identification of alloy compositions that satisfy user-defined performance targets.

2.2 Benchmarking on Auto-Data Pipeline

To construct a high-quality materials database, we develop a multi-step data extraction pipeline, termed **Auto-Data Pipeline**, which emphasizes structured task decomposition rather than single-step end-to-end (End2End) prompting. As shown in Figure 2a, the pipeline integrates literature filtering, attribute identification and extraction, cross-section information integration, and attribute standardization within a unified workflow. Moreover, the Auto-Data Pipeline constructs a structured, literature-derived database that systematically integrates alloy compositions, processing routes, testing types, and experimental conditions. Each reported measurement is explicitly linked to its corresponding testing methodology and conditions, forming complete composition–processing–testing–property records.

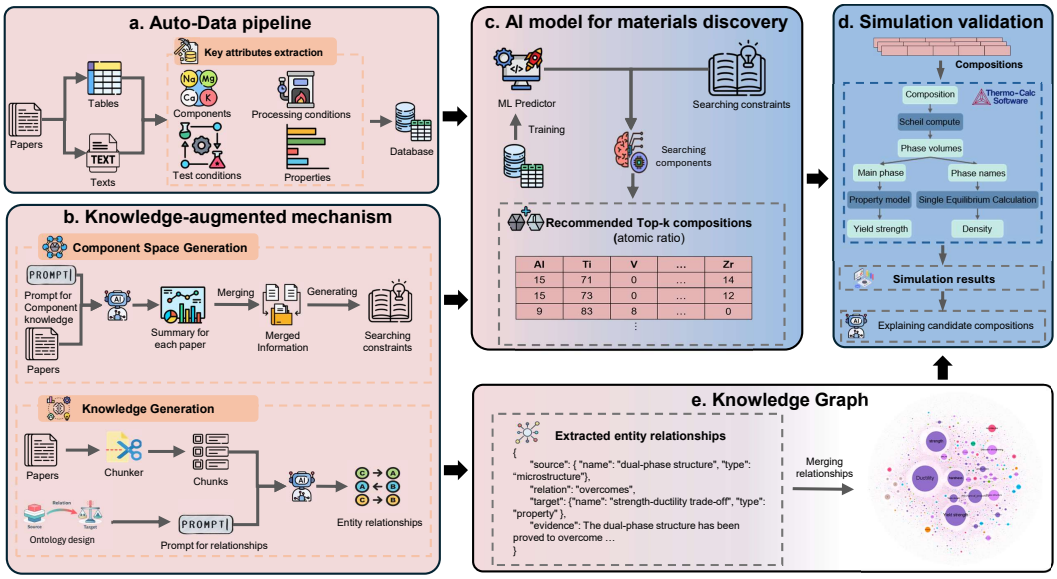


Figure 1: **Overall workflow of MatSeek**: a, a data pipeline extracts key attributes (Composition, processing, testing conditions, and properties) from papers into a structured database. b, a knowledge-augmented mechanism generates a recommended component space and extracts entity relations for knowledge construction. c, an AI predictor is trained with the extracted database and searches within the generated component space to recommend top-k compositions. d, thermodynamic simulations (e.g., Thermo-Calc) validate the recommended compositions and produce simulation results. e, extracted knowledge relations are merged into a unified knowledge graph to support interpretation and downstream discovery for human experts.

To rigorously assess the performance of Auto-Data Pipeline, we manually construct a gold-standard dataset. Extraction performance is assessed in terms of precision and recall, as summarized in Figure 2b. The results show that our pipeline consistently outperforms an end-to-end (End2End) prompting baseline (Giorgi et al., 2022; Zheng et al., 2023). In particular, when integrated with our pipeline, GPT-5 achieves a precision of 96.49% and a recall of 83.47% for yield strength, representing the balanced overall performance among the evaluated LLMs. While End2End attains comparable recall, it exhibits substantially lower precision due to spurious extractions. These results indicate that End2End prompting strategies struggle to reliably extract complex, structured materials data. We further evaluate the computational cost of different language models, with the results summarized in Appendix A. The extraction performance of the leading LLMs is generally similar. Given its favorable balance between precision and cost efficiency, GPT-5 is chosen for large-scale data extraction, resulting in 27,438 structured records from 10,240 HEA publications.

2.3 Analysis of the Database

Here, we conduct a comprehensive analysis of the extracted database using our Auto-Data Pipeline, which covers 10,240 papers and yields a total of 27,438 records. Figure 3a presents the temporal distribution of publications. From 2000 to 2025, the annual number of publications shows a strong overall upward trend, with rapid growth after 2011. HEAs consistently dominate the literature, whereas other alloys emerge more frequently in recent years, reflecting increasing diversification.

We further analyze the distributions of ultimate tensile strength (UTS, σ_y) and yield strength (YS, σ_β) under tensile testing, shown in Figure 3b. HEAs demonstrate the broadest strength–density envelope across all material classes, covering a wide density spectrum and achieving comparable yield and tensile strengths. Moreover, low-density Al- and Mg-based alloys cluster at low strengths, whereas Ni/Fe/Co-rich alloys populate the high-density region with high strengths. Importantly, the maximum specific strength (strength/density) points in the plots are attained by HEAs, indicating that HEAs can access favorable combinations of high strength and reduced density compared with conventional alloy families. The Sankey diagram in Figure 3c shows the distribution of yield strength across different alloy systems. High-entropy alloys dominate the dataset and span a wide

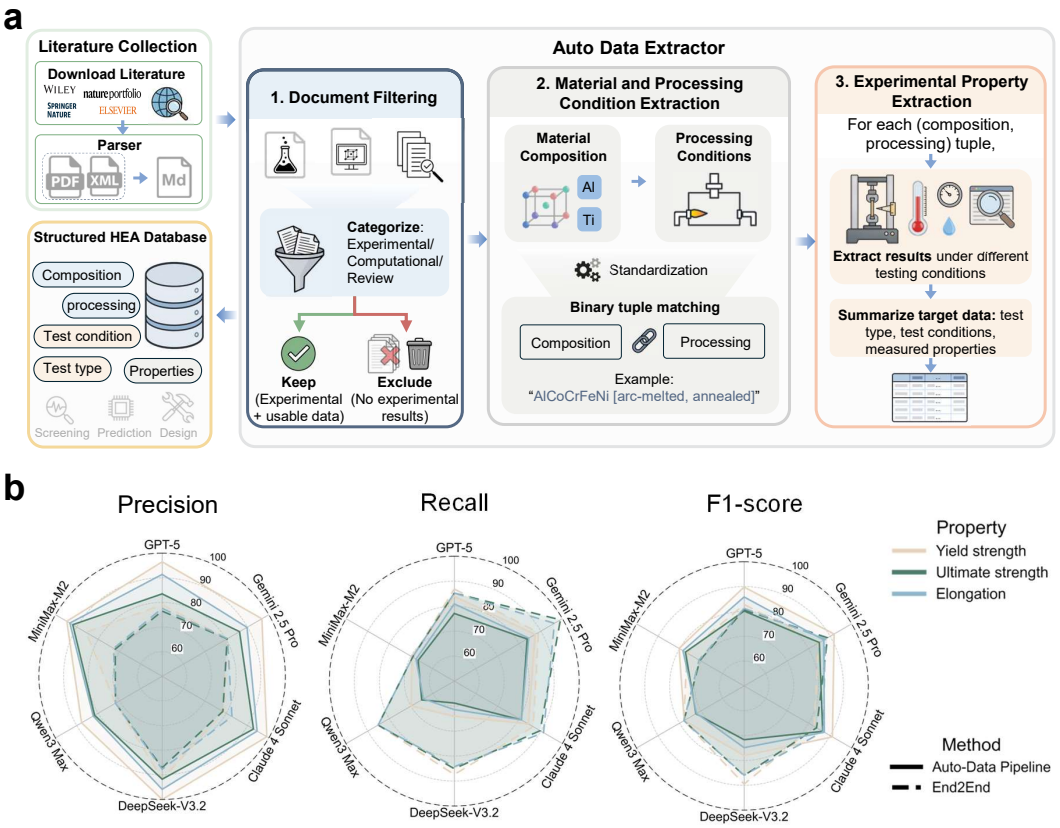


Figure 2: **Illustration of our Auto-Data pipeline and performance evaluation.** (a) The workflow of Auto-Data Pipeline. (b) These figures illustrate the performance of Auto-Data Pipeline on the extracted data records. We evaluate the common LLMs, including closed-sourced and open-sourced.

range of strength levels, with a noticeable fraction reaching the high-strength regime (>1000 MPa). In contrast, conventional alloys such as Fe-, Al-, and Mg-based systems are mainly concentrated in the low to medium yield strength categories.

Finally, we construct a knowledge graph based on the publications, where nodes represent material-related concepts extracted from text and edges indicate their co-occurrence relationships. We design an ontology (see Figure 8,) comprising eight entity types: composition, process, condition, property, microstructure, compound, element, and others. This ontology provides a unified schema to organize heterogeneous knowledge across studies. Node size reflects the relative frequency or salience of each concept in the literature, while node color encodes its ontology category. From the graph 3d, it is evident that the knowledge structure of the HEA literature is strongly centered on mechanical properties. Strength- and ductility-related concepts form the dominant hubs, reflecting the prevailing focus on the strength–ductility trade-off, while mechanistic and microstructural concepts remain secondary and mainly serve explanatory roles. Overall, the ontology-guided knowledge graph offers a structured representation of how different materials concepts are connected, forming the foundation for subsequent analysis of knowledge organization within the high-entropy alloy domain.

3 Knowledge-Assisted Alloy Design

Here, we deploy the extracted dataset on an alloy property prediction task. Specifically, we utilize the MATAI framework (Deng et al., 2025) based on the extracted dataset to train a deep neural network (DNN) predicting YS and UTS from alloy compositions. Then, we evaluate our model performance on the MPEA dataset (Borg et al., 2020). To ensure a fair comparison with this benchmark, we align the training by filtering and cleaning the data records to retain only as-cast tensile-test measurements (Figure 4a). Finally, we obtain 1,770 unique training records.

Chronological expansion improves predictive accuracy. Our results show that expanding the training dataset with newly extracted literature records steadily improves prediction performance on

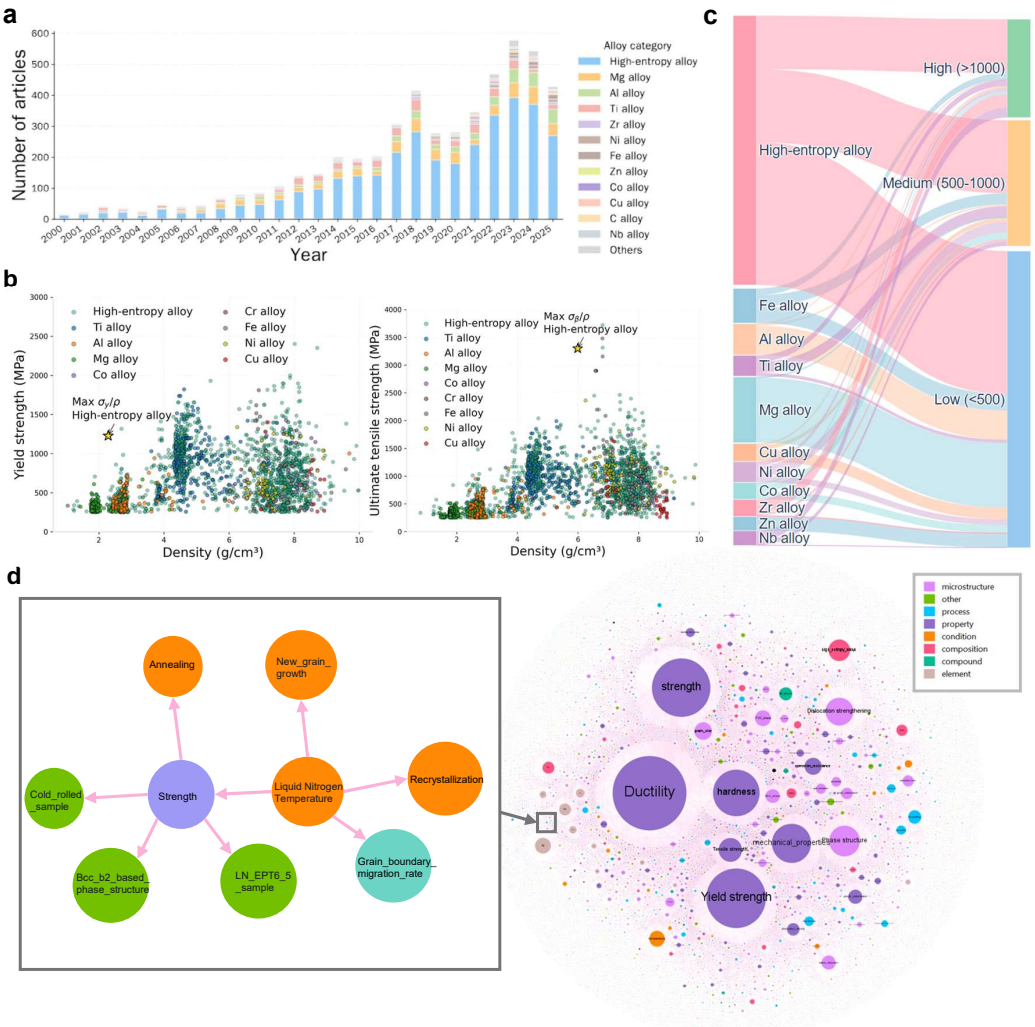


Figure 3: **Illustration of our Auto-Data pipeline and performance evaluation.** (a) The workflow of Auto-Data Pipeline. (b) These figures illustrate the performance of Auto-Data Pipeline on the extracted data records. We evaluate the common LLMs, including closed-sourced and open-sourced.

the MPEA benchmark. To understand this trend, we visualize the evolution of the alloy composition space using two-dimensional t-SNE embeddings (Figure 4b). Early training data (1997–2018) cover only a limited region of the composition space, whereas incorporating more recent studies progressively broadens the distribution. In the latest stage, test alloys are embedded within the major compositional clusters of the training set, indicating substantially improved coverage. Consistently, we can observe that predictive accuracy for both YS and UTS improves systematically as the dataset expands (Figure 4c).

Knowledge-Guided Design Spaces Enable More Efficient Inverse Alloy Optimization. We integrate literature-derived knowledge into inverse alloy design by using it to parameterize and constrain the composition search space, and benchmark the resulting spaces on the task of identifying HEAs with maximal specific yield strength (SYS, YS/Density). The design space in MATAI is specified as a JSON schema defining allowable elements, compositional bounds, and hard constraints. Given a design space, we run an iterative local-search optimizer guided by a DNN-based predictor trained on the full dataset (1997–2027), and evaluate candidate alloys using Thermo-Calc (Andersson et al., 2002) (version 2024b). We compare three specifications: (i) a vanilla space including all 27 TCHEA elements[1] without any constraints, (ii) an LLM-only space generated without literature grounding, and (iii) a literature-grounded MatSeek space constructed from extracted relational knowledge. Figure 4d shows that the best SYS as validated by Thermo-Calc, at each search step. The vanilla composition space shows slow improvement due to inefficient exploration over a large domain, while the

[1]Thermo-Calc supports 27 elements in the TCHEA database (TCHEA8).

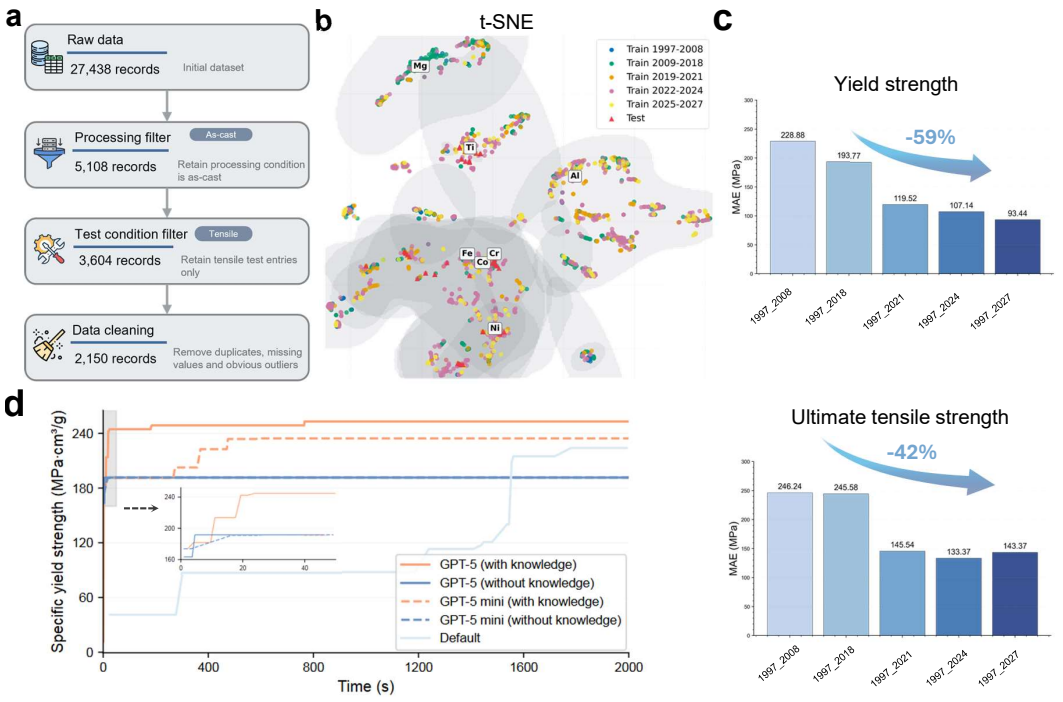


Figure 4: **Validation of the DNN-based predictors and semantic knowledge-guided inverse design.** a, Schematic illustration of the data cleaning and filtering procedure used to construct the training dataset for the DNN-based predictor. b, Two-dimensional t-SNE embedding of all alloy compositions, colored by publication year, illustrating the temporal expansion of composition space. c, Prediction performance of the DNN-based predictors as a function of the progressively expanded training datasets. d, Optimization trajectories for maximizing specific yield strength.

LLM-only space achieves rapid initial gains but quickly stalls, suggesting overly restrictive or misspecified bounds. In contrast, the composition space generated by the knowledge-enhanced LLM focuses exploration on high-value regions, leading to sustained improvement and consistently superior final performance. Notably, the best candidate identified by MatSeek is $Al26Ti44Zr_{30}$, which achieves the highest specific yield strength among all evaluated settings. Finally, we employ the knowledge-graph-based LLM to provide an interpretable explanation for the optimal composition. Specifically, GPT-5 offers the following mechanistic insights, as summarized below:

1. The addition of Al to Ti-based alloys significantly enhances strength while maintaining a low density. Notably, many Ti compositions with high Al content achieve top-ranked SYS.
2. Substituting Ti with Zr, or introducing Zr into Ti-based systems, can further improve strength with only a modest increase in density.
3. Non-equiatomic and Al-rich compositions often promote the formation of dual-phase microstructures or strengthening phases, which contribute to exceptionally high yield strength and, consequently, very high SYS.

4 Discussion

This work shows that effective knowledge-driven alloy discovery requires more than accurate predictors, it also depends on how knowledge from the literature is organized and made actionable. MatSeek not only automates large-scale extraction of structured alloy records, but also captures literature-derived scientific relations and converts them into usable priors and constraints for inverse design. By grounding the composition design space in this knowledge, MatSeek avoids both the inefficiency of unconstrained searches and the early stagnation of overly restrictive, ungrounded spaces, leading to consistently improved optimization trajectories. These results highlight that the literature is not just a source of reported measurements, but also a reusable repository of expert design knowledge for inverse materials design.

Our current implementation focuses on single-objective optimization, while practical alloy design often involves multi-property trade-offs. Future work will extend MatSeek to multi-objective decision-making and incorporate richer processing- and microstructure-aware models, as well as experimental or simulation feedback to further close the loop.

4.1 AUTO-DATA PIPELINE

The Auto-Data Pipeline is designed to construct a structured, literature-derived materials database by systematically extracting and normalizing composition–processing–property information from heterogeneous scientific publications. The pipeline emphasizes document-level context integration and standardized data representation to ensure the reliability and comparability of extracted records.

Scientific full-text publications are collected from online sources in PDF or XML formats and converted into structured plain-text representations prior to extraction. During this preprocessing stage, textual content and tables are preserved, while figures are omitted, as quantitative materials information is consistently reported in the text or tabulated data. This unified text representation enables robust downstream interpretation by LLMs. The Auto-Data Pipeline consists of three stages (see Figure 2). First, literature filtering is applied to exclude publications that lack relevant experimental results, thereby reducing noise introduced by keyword-based retrieval and improving extraction efficiency. Second, alloy compositions and their associated processing procedures are jointly extracted and standardized, ensuring explicit linkage between chemical composition and fabrication history within each material record. Third, for each composition–processing pair, all reported target properties are extracted together with their corresponding testing methodologies and experimental conditions, enabling the construction of well-defined and internally consistent performance records.

Evaluation Metrics We evaluate the performance of the Auto-Data Pipeline using precision, recall, and F1 score. Each material record is uniquely defined by a tuple of two attributes: alloy composition and test type. Physical property values are compared only between records whose attribute tuples are determined to be equivalent.

$$\text{Precision} := \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \quad \text{Recall} := \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Many publications report multiple compositions but omit mechanical properties for some of them, resulting in missing values. During evaluation, we handle such cases as follows. If both the extracted value and the ground-truth value are non-empty and match, the case is counted as a true positive. If the extracted value is non-empty but the ground-truth value is either empty or does not match, the case is counted as a false positive. If the ground-truth value is non-empty but the extracted value is empty, the case is counted as a false negative. Cases in which both the extracted and ground-truth values are empty are excluded from evaluation, as they do not convey meaningful information. Under this definition, the number of non-empty extracted values equals the sum of true positives and false positives. Precision, recall, and F1 score are then computed based on these counts.

4.2 ALLOY PROPERTY PREDICTION

In this work, we train the predictor using only as-cast tensile records. Starting from 27,438 extracted data entries, we filter to 3,604 as-cast measurements. For duplicate compositions, the minimum value is retained. After cleaning and outlier removal, the final training dataset contains 1,770 records (Figure 4a). Then, the DNN-based predictor of MatSeek is built on the MATAI (Deng et al., 2025), which provides general-purpose, physics-aware models for alloy property prediction and inverse design. Alloy compositions are represented as elemental atomic-fraction vectors and used as inputs to train predictors for key mechanical properties, including yield strength and ultimate tensile strength. Multi-task learning is employed to jointly predict YS and UTS, with an explicit physical constraint, i.e., UTS $\geq$ YS, thereby improving model robustness and physical consistency. The trained predictors serve as fast surrogate models for property evaluation and are coupled with a constraint-aware inverse design engine that formulates alloy discovery as a discrete constrained optimization problem. To efficiently explore large compositional spaces under hard physical and manufacturability constraints, MatSeek adopts a bi-level optimization strategy that combines stochastic local search with symbolic constraint programming. Importantly, the semantic relational knowledge extracted

from the literature is directly integrated into the definition of design variables, domains, and constraints, enabling knowledge-guided exploration of composition space and efficient identification of high-performance alloy candidates.

4.3 Semantic Knowledge-assisted Mechanism

Knowledge-guided alloy design space. To incorporate semantic knowledge into ML-based inverse alloy design, we develop a knowledge-guided workflow to construct SYS-oriented alloy design constraints. The framework consists of two stages. In the first stage, LLMs extract relevant information from each publication, including alloy compositions, processing routes, and microstructural features associated with SYS and yield strength. In the second stage, LLMs leverage the extracted knowledge to identify composition design principles, distinguish beneficial from detrimental alloying elements, and infer element combinations frequently linked to improved SYS. Based on these insights, we construct an SYS-focused design space that provides composition-level guidance for subsequent inverse optimization.

Knowledge graph generation. To construct a literature-derived knowledge graph (Dagdelen et al., 2024; Yang et al., 2022; Bai et al., 2025) from heterogeneous HEA publications, we employ a three-stage pipeline that converts raw Markdown documents into a unified graph representation. First, we remove non-semantic artifacts, such as references, figure captions, and formatting noise, in each document and then the cleaned text is segmented into overlapping sentence-level chunks using a sliding-window strategy, which preserves relations spanning adjacent sentences while maintaining local semantic coherence. Next, relation extraction is then performed exhaustively over all chunks. Knowledge extraction is restricted to explicit, text-supported relations, with entities limited to a predefined set of node types (see Figure 8). In this work, we use DeepSeek V3.1 (Liu et al., 2024) to extract the relations because of its low computational cost. In the final stage, the extracted relations are unified through a canonicalization and graph consolidation pipeline to produce a semantically consistent global knowledge graph. Because chunk-level extraction inevitably introduces surface-form variations (e.g., "strengthening effect," "enhanced tensile strength," and "UTS"), we apply a multilayer normalization strategy that combines rule-based cleaning, domain-specific dictionaries, and embedding-based similarity matching using SentenceTransformer (Reimers & Gurevych, 2019) embeddings. After deduplication and noise filtering, semantically equivalent entities are merged into stable canonical nodes, and the resulting triples are integrated into a global graph. Although relations are extracted independently at the chunk level, shared canonical entities enable higher-order relational paths to emerge across documents, yielding a coherent representation of HEAs that supports downstream retrieval, visualization, and scientific reasoning.

REFERENCES

Jan-Olof Andersson, Thomas Helander, Lars Höglund, Pingfang Shi, and Bo Sundman. Thermo-calc & dictra, computational tools for materials science. *Calphad*, 26(2):273–312, 2002.

Xuefeng Bai, Song He, Yi Li, Yabo Xie, Xin Zhang, Wenli Du, and Jian-Rong Li. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

Christopher KH Borg, Carolina Frey, Jasper Moh, Tresa M Pollock, Stéphane Gorsse, Daniel B Miracle, Oleg N Senkov, Bryce Meredig, and James E Saal. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Scientific Data*, 7(1):430, 2020.

Mouyang Cheng, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. Artificial intelligence-driven approaches for materials design and discovery. *Nature Materials*, pp. 1–17, 2026.

Jaewoong Choi and Byungju Lee. Accelerating materials language processing with large language models. *Communications Materials*, 5(1):13, 2024.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418, 2024.

Yanchen Deng, Chendong Zhao, Yixuan Li, Bijun Tang, Xinrun Wang, Zhonghan Zhang, Yuhao Lu, Penghui Yang, Jianguo Huang, Yushan Xiao, et al. Matai: A generalist machine learning framework for property prediction and inverse design of advanced alloys. *arXiv preprint arXiv:2511.10108*, 2025.

Ali Riza Durmaz, Akhil Thomas, Lokesh Mishra, Rachana Niranjan Murthy, and Thomas Straub. An ontology-based text mining dataset for extraction of process-structure-property entities. *Scientific data*, 11(1):1112, 2024.

Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 4(1):2356506, 2024.

John Giorgi, Gary Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (eds.), *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 10–25, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.2. URL https://aclanthology.org/2022.bionlp-1.2/.

Stéphane Gorsse, MH Nguyen, Oleg N Senkov, and Daniel B Miracle. Database on the mechanical properties of high entropy alloys and complex concentrated alloys. *Data in brief*, 21:2664–2678, 2018.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

Liuliu Han, Shuya Zhu, Ziyuan Rao, Christina Scheu, Dirk Ponge, Alfred Ludwig, Hongbin Zhang, Oliver Gutfleisch, Horst Hahn, Zhiming Li, et al. Multifunctional high-entropy materials. *Nature Reviews Materials*, 9(12):846–865, 2024.

Shuyuan Li, Yunjiang Zhang, Zhaolin Fang, Kong Meng, Rui Tian, Hong He, and Shaorui Sun. Extracting the synthetic route of pd-based catalysts in methanol steam reforming from the scientific literature. *Journal of Chemical Information and Modeling*, 63(20):6249–6260, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Radeen Mostafa, Mirza Nihal Baig, Mashaekh Tausif Ehsan, and Jakir Hasan. G-rag: Knowledge expansion in material science. *arXiv preprint arXiv:2411.14592*, 2024.

Zongrui Pei, Junqi Yin, Peter K Liaw, and Dierk Raabe. Toward the design of ultrahigh-entropy alloys via mining six million texts. *nature communications*, 14(1):54, 2023.

Zongrui Pei, Junqi Yin, Jörg Neugebauer, and Anubhav Jain. Towards the holistic design of alloys with large language models. *Nature Reviews Materials*, 9(12):840–841, 2024.

Olivia P Pfeiffer, Haihao Liu, Luca Montanelli, Marat I Latypov, Fatih G Sen, Vishwanath Hegadekatte, Elsa A Olivetti, and Eric R Homer. Aluminum alloy compositions and properties extracted from a corpus of scientific manuscripts and us patents. *Scientific Data*, 9(1):128, 2022.

Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.

Tresa M Pollock. Alloy design for aircraft engines. *Nature materials*, 15(8):809–815, 2016.

Thomas Michael Pruyn, Amro Aswad, Sartaaj Takrim Khan, Ju Huang, Robert Black, and Seyed Mohamad Moosavi. Mof-chemunity: Literature-informed large language models for metal–organic framework research. *Journal of the American Chemical Society*, 147(47):43474–43486, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

JM Rickman, HM Chan, MP Harmer, JA Smeltzer, CJ Marvel, A Roy, and G Balasubramanian. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nature communications*, 10(1):2618, 2019.

Matthew C Swain and Jacqueline M Cole. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904, 2016.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Zheren Wang, Olga Kononova, Kevin Cruse, Tanjin He, Haoyan Huo, Yuxing Fei, Yan Zeng, Yingzhi Sun, Zijian Cai, Wenhao Sun, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific data*, 9(1):231, 2022.

Rongen Yan, Xue Jiang, Weiren Wang, Depeng Dang, and Yanjing Su. Materials information extraction via automatically generated corpus. *Scientific Data*, 9(1):401, 2022.

Penghui Yang, Chendong Zhao, Bijun Tang, Zhonghan Zhang, Xinrun Wang, Yanchen Deng, Yuhao Lu, Cuntai Guan, Zheng Liu, and Bo An. Automat: A hierarchical framework for autonomous alloy discovery. *arXiv preprint arXiv:2507.16005*, 2025.

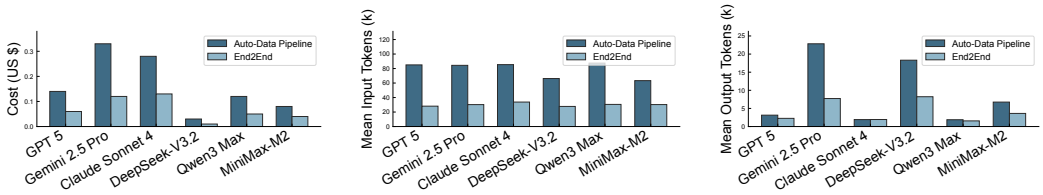
Xianjun Yang, Ya Zhuo, Julia Zuo, Xinlu Zhang, Stephen Wilson, and Linda Petzold. Pcmsp: A dataset for scientific action graphs extraction from polycrystalline materials synthesis procedure text. *arXiv preprint arXiv:2210.12401*, 2022.

Zian Zhang, Haoxuan Tang, and Zhiping Xu. Fatigue database of complex metallic alloys. *Scientific Data*, 10(1):447, 2023.

Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023. doi: 10.1021/jacs.3c05819. PMID: 37548379.

A Cost of Auto-Data Pipeline

Figure 5 compares the cost of Auto-Data Pipeline and End2End across different LLMs. Although Auto-Data Pipeline incurs a higher cost, it achieves substantially higher precision, as shown in Figure 2b.



(a) Cost (b) Input tokens (c) Output tokens

Figure 5: **Cost and token usage comparison across LLMs.** Auto-Data Pipeline vs End2End in terms of (a) cost, (b) input tokens, and (c) output tokens.

B Analysis of the extracted database

Figure 6a reveals a clear strength–ductility trade-off across different alloy systems, with high-entropy alloys exhibiting the broadest performance range and achieving superior combinations of high ultimate tensile strength and elongation.

Figure 6b further reveal strong co-occurrence patterns among key alloying elements and frequently adopted processing routes, with casting and heat treatment serving as dominant pathways, highlighting common design and manufacturing strategies in high-performance alloy development.

The left Sankey diagram in Figure 6c illustrates the temporal distribution of processing routes reported in the literature. Casting remains the most frequently adopted method across all years, while advanced treatments such as heat treatment, quenching, and rolling appear increasingly in more recent publications, reflecting a growing focus on process optimization for enhanced alloy performance. The right Sankey diagram in Figure 6c shows the relationship between processing routes and yield strength categories. Conventional routes such as casting and arc melting are mainly associated with low yield strength ($<$500 MPa), whereas strengthening processes including heat treatment, quenching, and intensive deformation contribute more substantially to the medium and high strength regimes, highlighting the importance of thermo-mechanical processing in achieving superior mechanical properties.

C Results of DNN-based predictors

In this section, we report the R^2 scores of alloy property predictors trained on chronologically expanded datasets. As shown in Figure 7, predictive performance improves substantially as newly extracted literature records are progressively incorporated into the training set. For yield strength, the model accuracy increases from near-zero performance in the earliest stage to an R^2 of 0.757 when trained on the full dataset up to 2027, corresponding to an overall improvement of more than 3700%. A similar trend is observed for ultimate tensile strength, where the R^2 score rises to 0.619 with an improvement exceeding 2600%. These results highlight that continued literature-driven dataset expansion systematically enhances model generalization on the MPEA benchmark.

D More details of MatSeek

Ontology design of the knowledge graph To effectively construct the knowledge graph, we design a domain-specific ontology to extract key entities from the literature. As shown in Figure 8, the ontology defines different node types, including elements, compounds, compositions, processing routes, testing conditions, microstructures, and material properties, as well as their relationships.

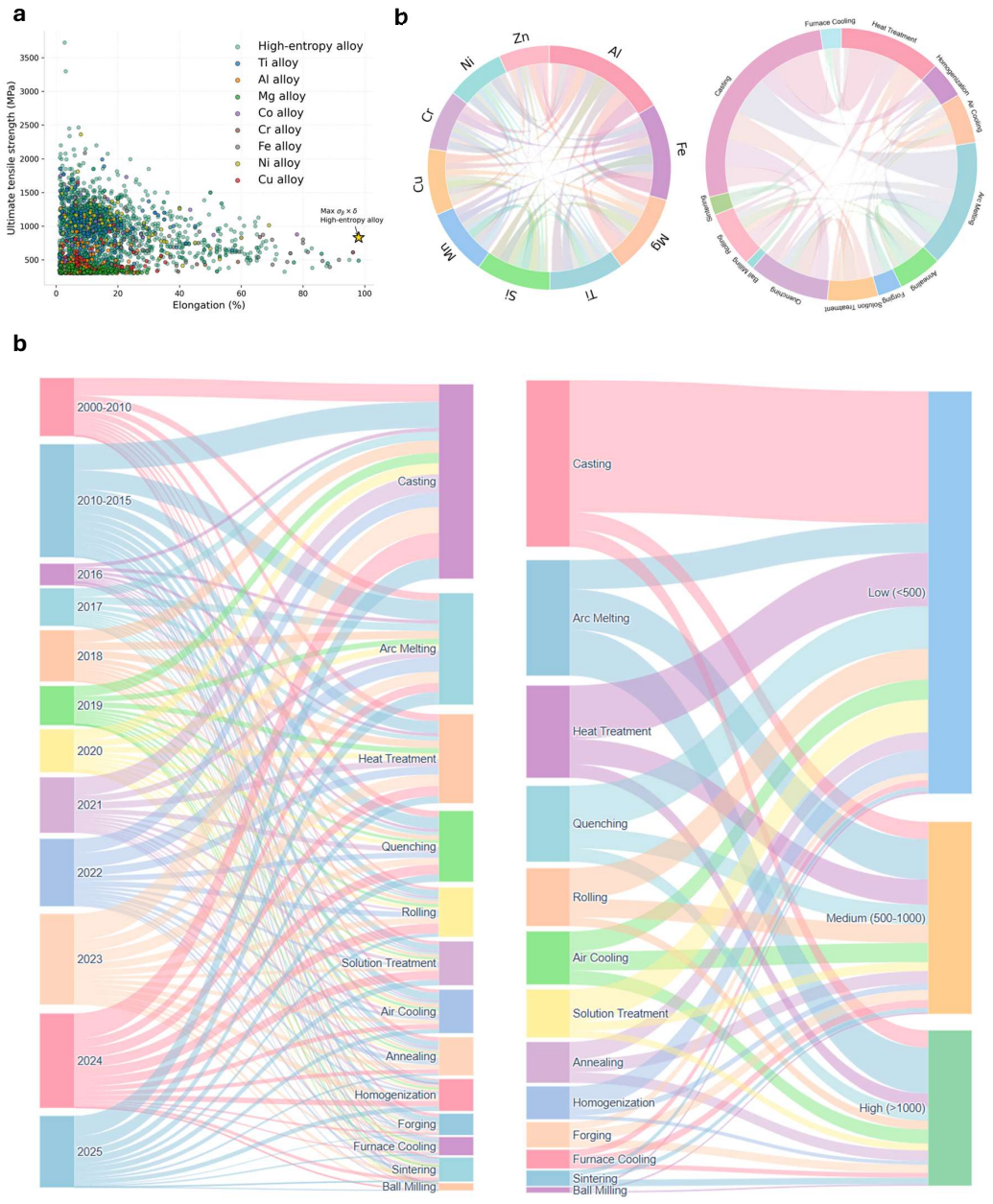


Figure 6: **More data analysis of the extracted data.**

This structured representation enables systematic knowledge extraction and supports downstream materials discovery and interpretation.

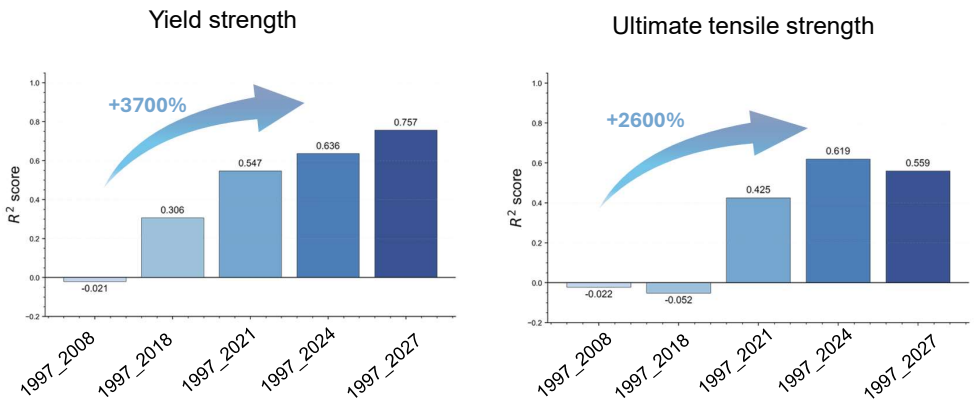


Figure 7: R^2 **scores of different predictors trained on chronologically expanded datasets.**

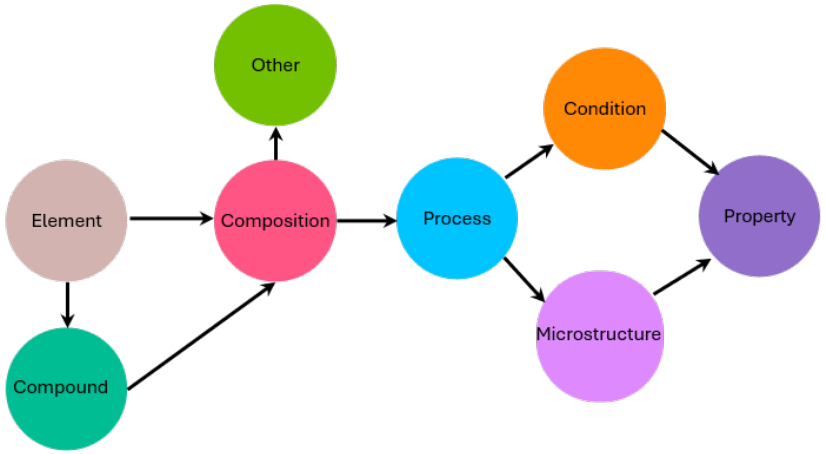


Figure 8: The ontology design of the knowledge mining.