MatSeek: An Automated Knowledge-Driven Framework for Materials Research

Jianguo Huang [1] * **Xuyu Dong** [2 3] * **Yuhao Lu** [1] **Yanchen Deng** [1] **Chendong Zhao** [2] **Penghui Yang** [1] **Zhonghan Zhang** [2] **Yixuan Li** [1] **Yushan Xiao** [2] **Cuntai Guan** [1] **Bijun Tang** [2] **Xinrun Wang** [4] **Zheng Liu** [2 5] **Bo An** [1]

*Equal contribution [1]*College of Computing and Data Science, Nanyang Technological University, Singapore* [2]*School of Materials Science and Engineering, Nanyang Technological University, Singapore* [3]*School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China* [4]*School of Computing and Information Systems, Singapore Management University, Singapore* [5]*CINTRA CNRS/NTU/THALES, UMI 3288, Research Techno Plaza, 50 Nanyang Drive, Border X Block, Level 6, Singapore 637553, Singapore*. Correspondence to: bjtang@ntu.edu.sg, xrwang@smu.edu.sg, ctguan@ntu.edu.sg, z.liu@ntu.edu.sg, boan@ntu.edu.sg.

1. Introduction

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. Further experiments verified the lightweight alloy that breaks through the existing Pareto frontier. MatSeek establishes a unified, literature-grounded paradigm for knowledge-driven materials discovery.

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.

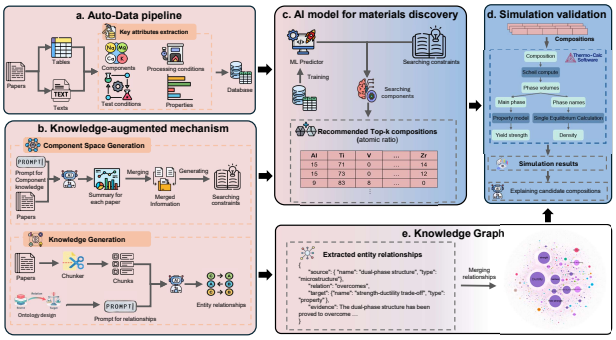

Fig. 1: **Overall workflow of MatSeek**: a, a data pipeline extracts key attributes (Composition, processing, testing conditions, and properties) from papers into a structured database. b, a knowledge-augmented mechanism generates a recommended component space and extracts entity relations for knowledge 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 validate the reliability of the ML model predictions, the resulting candidate compositions are assessed using thermodynamic simulation software, such as Thermo-Calc [1]. 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 in-

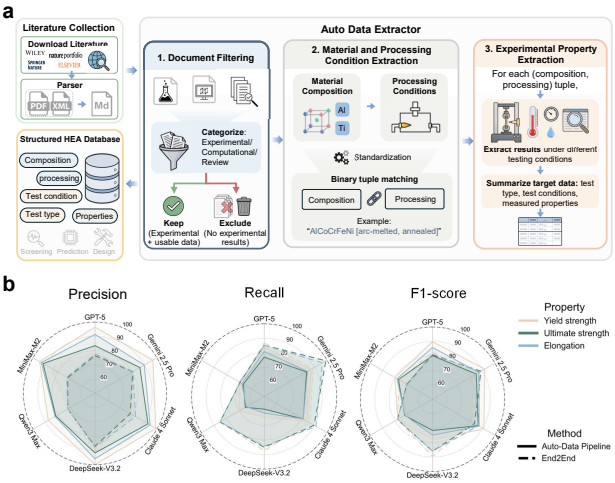

Fig. 2: **Illustration of our Auto-Data pipeline and performance evaluation.** (a) The workflow of Auto-Data Pipeline. (b) These figures illustrate the performance of Auto-Data Pipeline on the extracted data records. We evaluate the common LLMs, including closed-sourced and open-sourced.

formation 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.

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 [2, 3]. 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. 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 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.

References

[1] 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.

[2] 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, editors, *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] 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. PMID: 37548379.