MATERIALS RESEARCH AGENT

Mohd Zaki
Hopkins Extreme Materials Institute
Johns Hopkins University
Baltimore, MD, USA
mzaki4@jhu.edu

Michael D. Shields
Hopkins Extreme Materials Institute
Department of Civil and Systems
Johns Hopkins University
Baltimore, MD, USA
michael.shields@jhu.edu

ABSTRACT

Materials discovery pipelines rely on the availability of high-quality datasets, accurate predictive models, and exploration tools for searching through published experimental and simulation-based findings. The number of documents containing valuable data (e.g., peer-reviewed research papers, patents, and handbooks) is quite high, making manual compilation of datasets a challenging task. Further, developing reliable machine learning models for predicting material properties with uncertainty estimates and explanations is highly desirable for rational experimental planning. Therefore, we propose an agentic approach for materials discovery that combines (a) information extraction from research papers to create machine-readable materials composition–property datasets, (b) training uncertainty-aware property prediction models, (c) delineating the effect of input features on material properties through explainable AI techniques, and (d) materials selection charts to assist in identifying potential compositions of interest required to push existing material–property frontiers.

1 INTRODUCTION

Developing materials for target applications is a challenging task (Jain et al. (2020); Venugopal et al. (2021)). First, existing materials data are scattered across structured (tables) and unstructured (text and images) parts of scientific documents, making search difficult (Venugopal et al. (2021); Venugopal & Olivetti (2024); Olivetti et al. (2020)). Further, the tools required to extract meaningful insights from a large number of scientific documents are either non-interoperable or difficult to use. To address these issues, several researchers have proposed using large language models (LLMs) to facilitate information extraction (IE) with minimal or no supervision (Dagdelen et al. (2024); Lee et al. (2025); Yuan et al. (2026)).

Considering the significant computational resources required to develop LLMs for IE at large scale, Hira et al. (2025) adapted a hybrid pipeline of materials-science domain-specific IE rules, graph neural networks, and small language models (Gupta et al. (2023; 2022)) to create accurate, high-speed composition and property extraction tools. Further, to obtain actionable insights from the output of IE tools, understanding the effect of composition, structure, processing and testing methods on materials properties is highly important (Zaki et al. (2022); Olivier et al. (2021)). Additionally, this understanding should also encompass the effect of uncertainties to guide rational experimentation and simulation campaigns for targeted materials development. Hence, an easy-to-use, interactive, open-source tool that is capable of extracting information from a large number of research papers, understanding composition–property relationships and associated uncertainties, would be beneficial for materials science researchers. To this end, we propose the Materials Research Agent (MATRA), which is a collection of state-of-the-art tools for IE and uncertainty quantification.

To demonstrate the capabilities of MATRA, we use it to create a database of materials used in extreme environments (Fahrenholtz & Hilmas (2017)). The database is then used to develop Bayesian neural networks to model composition–property relationships, followed by visualizing uncertainty-aware materials selection charts (MSCs). In the following sections, we provide a brief description of all tools, followed by results on corpus development, ML model training, prediction explanations, and MSC construction. We also discuss limitations of the proposed system and directions for future work to improve the functionality of MATRA.

2 METHODOLOGY

MatRA comprises specialized tools for each task, as shown in Fig. 1. The GitHub repository provides further details about MATRA. The description of each tool is as follows:

2.1 CORPUS PREPARATION

The capability of AI agents to drive autonomous materials research depends on underlying databases. In this section, we describe how different tools are used to prepare the corpus according to user inputs. Note that the access to research papers that can be used to create the database depends on the institutional/user subscriptions.

2.1.1 QUERY SEARCH

This tool utilizes the CrossRef API to retrieve research paper DOIs according to search queries. The list of keywords for creating search queries can either be provided by the user, or the agent generates it based on the natural language description provided as input. The API can be configured to retrieve research papers from a specific journal based on publication dates, author names, affiliations, etc. Readers can refer to the following GitHub repository to learn more about the CrossRef API.

2.1.2 XML DOWNLOADING

Once the paper DOIs are provided by the query search tool, they are now utilized by the research paper downloading tool to obtain the XML of each paper. The XML for research papers published after 1999 is available through the ScienceDirect API (els). This tool first creates the folder for each journal, and then retrieves the XML using the API. The XML is parsed to obtain the PII of the paper which is then used to name the subfolder where the XML is stored. Overall, this tool stores the paper XML in sub-folders of respective journal folders. The sub-folders are named using the PII and the XML is also stored with the same name inside respective sub-folders.

2.1.3 MATSKRAFT

Since IE using LLMs from a large number of research papers is computationally and economically prohibitive, we use the state-of-the-art materials-science-specific framework MATSKRAFT (Hira et al. (2025)). It combines a materials science language model (MatSciBERT; Gupta et al. (2022)), graph neural network (GNN) models, and rule-based systems to extract and store material compositions and properties in machine-readable format. It takes as input tables and text from materials science research papers. Specifically, the XML files downloaded in the previous step are used as input to this tool. First, the text is extracred from different paper sections and stored in JSON format. Next, the tables are extracted from each paper and stored as a list of dictionaries, where each dictionary uses the PII as a key and stores the list of tables. Tables are stored in a list-of-lists format, where each row is converted to a list. MATSKRAFT then processes these data to compile material compositions and properties. Note that the preparation of input files for MATSKRAFT can be done in parallel, depending on the number of available CPU cores. The GNNs in MATSKRAFT can be run on an NVIDIA V100 32GB GPU.

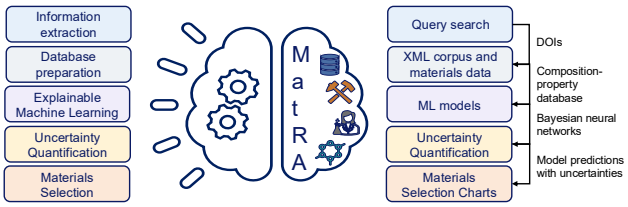

Figure 1: Functionalities available in MATRA and its application

2.2 MACHINE LEARNING MODEL DEVELOPMENT AND EXPLAINABILITY

Rational experimental planning requires information about uncertainty in predicted values of material properties. Therefore, this tool is responsible for training a Bayesian neural network (BNN)

implemented using the UQpy library (Krill et al. (2025); Olivier et al. (2020)). The hyperparameters made available for tuning include the number of hidden layers, neurons in each layer, learning rate, regularization parameter (weight decay), beta, batch size, and number of epochs. Given the large hyperparameter optimization space, MATRA uses Optuna (Akiba et al. (2019)) to obtain the best ML model. Optuna provides a scalable define-by-run framework with efficient sampling and pruning strategies. A set of hyperparameters (trial) is treated as input, and the validation score is treated as output; Optuna attempts to maximize this score to obtain the best model. Since the BNN is based on PyTorch, MATRA uses the DeepExplainer algorithm in SHAP (Lundberg & Lee (2017); Shrikumar et al. (2017)). It approximates the conditional expectations of SHAP values by integrating over background samples. The resulting SHAP values for an input feature sum up to the difference between the expected model output on the passed background samples and the current model output.

2.3 MATERIAL SELECTION CHARTS

Functional materials for real-world applications are expected to satisfy multiple property constraints, which can be difficult to meet simultaneously. Therefore, MSCs, proposed by Ashby & Cebon (1993), are important for visualizing the existing Pareto front between desired material properties and planning experimental campaigns. This tool takes as input CSV files containing composition data and corresponding properties. It generates scatter plots that show clusters based on the elements present in the material compositions and plots material indices in the background. Further, the predicted material properties can be used to visualize uncertainty-aware MSCs, where the scatter points are associated with error bars for each property.

3 RESULTS AND DISCUSSION

The tools in MATRA can be used independently to perform each task, or they can be driven by LLMs. In this work, we demonstrate the application of MATRA to study materials used in extreme environments. The first step involves searching for related research papers through query-based search using the CrossRef API tool. Since carbides are a popular family of materials used in extreme environments, the search queries include the word *carbide* followed by symbols of elements from Group 4, 5, and 6 of the periodic table. The search returned millions of DOIs, but we considered only papers published in nineteen popular Elsevier journals in the domain. The XMLs of these papers were downloaded using the corpus preparation tool and provided as input to MATSKRAFT. These steps resulted in a database of 9,639 research papers, which contributed to the composition–property database. The elements present in the extracted compositions are shown in Fig. 2(a), and the frequency of each property in the extracted entries is shown in Fig. 2(b).

To demonstrate the functionality of the ML model development tool, we chose to predict density as a function of material composition. The extracted dataset is arranged into a CSV file, where each feature column represents an element. For a given row representing a material, if an element is present, its percentage is used as the feature value, while zero is assigned to features for elements that are absent. Fig. 2(c) shows actual versus predicted values from the trained ML models, followed by Fig. 2(d), which shows the effect of individual elements on the predicted values. In Fig. 2(d), the points in front of each element represent material compositions, and the point color indicates the relative abundance of that element in each material.

Fig. 2(e) shows a materials selection chart for identifying materials based on Young's modulus and density. The error bars associated with each point are obtained by sampling the parameters of the trained BNNs. The point color indicates the material family, defined by the names of the constituent elements.

4 LIMITATIONS AND FUTURE WORK

The current implementation of MATRA has been tested only in single-user, sequential mode. Since the tools can run independently when the required inputs are available, the agent can be extended to support multiple users and enable shared corpora and databases. Property extraction is currently limited to 18 properties, as shown in Fig. 2(b), due to the non-generative nature of the underlying

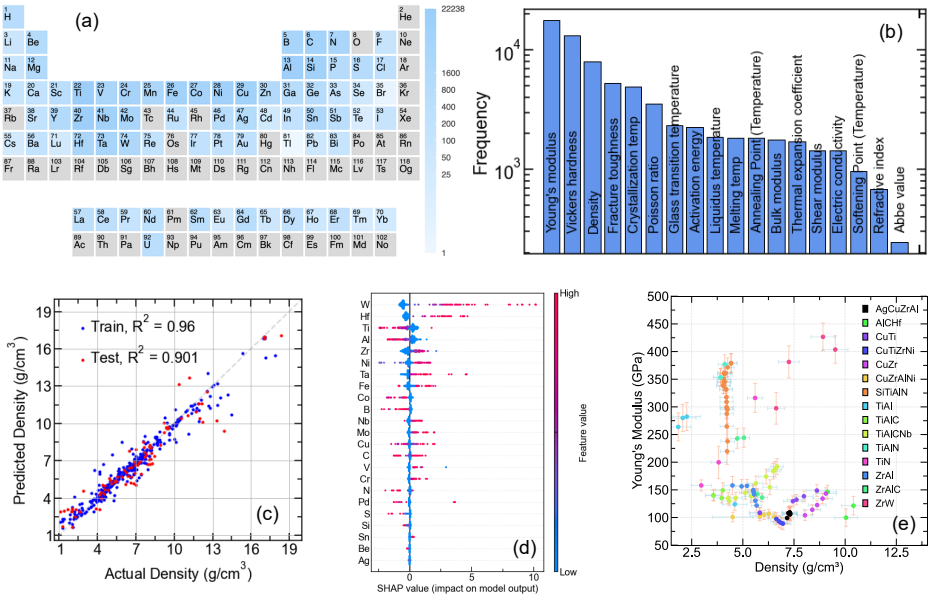

Figure 2: Visualizing the output of different tools available in MATRA

model. However, given its performance and computational advantages over LLM-based information extraction, there is a need to tailor it to extract additional properties.

Several open-source and proprietary AI agents can perform tasks essential for materials discovery. However, there is a lack of benchmarks to evaluate each step of the process, from information extraction to processing, synthesis, and testing. Further, agents should be evaluated based on their ability to plan tool use correctly and create new tools when required. Currently, we do not provide a facility to explicitly chat with the LLM driving MATRA, but this will be added in a future release after proper benchmarking.

REFERENCES

ScienceDirect.com | Science, health and medical journals, full text articles and books. URL https://www.sciencedirect.com/.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

Michael F Ashby and David Cebon. Materials selection in mechanical design. *Le Journal de Physique IV*, 3(C7):C7–1, 1993.

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, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45563-x.

William G Fahrenholtz and Greg E Hilmas. Ultra-high temperature ceramics: materials for extreme environments. *Scripta materialia*, 129:94–99, 2017.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1): 102, 2022.

Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan, and Mausam . DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.753. URL https://aclanthology.org/2023.acl-long.753.

Kausik Hira, Mohd Zaki, NM Krishnan, et al. Matskraft: A framework for large-scale materials knowledge extraction from scientific tables. *arXiv preprint arXiv:2509.10448*, 2025.

Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils ER Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pp. 1751–1784, 2020.

Connor Krill, Ponkrshnan Thiagarajan, George D Pasparakis, Somdatta Goswami, Dimitrios Tsapetis, Dimitris G Giovanis, and Michael D Shields. Uqpy version 4.2: Uncertainty quantification with python. *SoftwareX*, 32:102364, 2025.

Sanghoon Lee, Kevin Cruse, Viktoriia Baibakova, Gerbrand Ceder, and Anubhav Jain. Text-mined dataset of solid-state syntheses with impurity phases using large language model. *Scientific Data*, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317, December 2020. ISSN 1931-9401. doi: 10.1063/5.0021106.

Audrey Olivier, Dimitris G Giovanis, BS Aakash, Mohit Chauhan, Lohit Vandanapu, and Michael D Shields. Uqpy: A general purpose python package and development environment for uncertainty quantification. *Journal of Computational Science*, 47:101204, 2020.

Audrey Olivier, Michael D Shields, and Lori Graham-Brady. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Computer methods in applied mechanics and engineering*, 386:114079, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153. JMLR.org, 2017.

Vineeth Venugopal and Elsa Olivetti. MatKG: An autonomously generated knowledge graph in Material Science. *Scientific Data*, 11(1):217, February 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03039-z.

Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7), 2021.

Haolun Yuan, Jun Zeng, Jie Zuo, Xin Wang, and Dingguo Xu. A general llm-powered text mining framework: Applied to extract high entropy alloys. *Computational Materials Science*, 264: 114476, 2026.

Mohd Zaki, NM Anoop Krishnan, et al. Extracting processing and testing parameters from materials science literature for improved property prediction of glasses. *Chemical Engineering and Processing-Process Intensification*, 180:108607, 2022.