Mapping Materials Science: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature

Magdalena Lederbauer [1,2] **Siddharth Betala** [1] **Amine Sehaba** [3] **Alfonso Amayuelas** [4]
Anamaria Leonescu [1] **Valerie Genzke** [5] **Alexandre Duval** [1] **Mathilde L. D. Franckel** [1,6] **Samuel P. Gleason** [1]
[1]*Entalpic, France* [2]*Massachusetts Institute of Technology, USA* [3]*ENSA Lyon UMR-MAP Aria, France* [4]*University of California, Santa Barbara, USA* [5]*Independent Researcher, Germany* [6]*Imperial College London, UK*. Correspondence to: Magdalena Lederbauer magled@mit.edu, Samuel P. Gleason samuel.gleason@entalpic.ai.

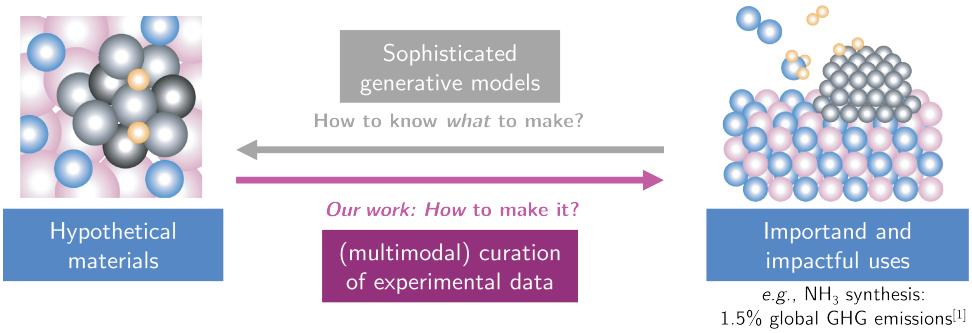

Fig. 1: Schematic overview of our multi-modal approach to materials discovery. Generative models can propose hypothetical materials (left), but realizing them experimentally requires synthesis knowledge currently locked in unstructured literature. Our framework curates multi-modal databases linking synthesis procedures to performance data—illustrated here for ammonia synthesis catalysts, where connecting preparation methods to catalytic activity enables data-driven optimization of this energy-intensive process responsible for 1.5% of global greenhouse gas emissions [1].

1. Introduction

The discovery of novel inorganic materials has proven essential for advancing technologies in energy conversion, storage, and catalysis [2, 3]. However, progress in materials discovery is limited by fragmented knowledge: synthesis protocols are scattered across decades of literature in unstructured formats, while performance metrics are predominantly presented as figures and plots. Breakthroughs in adjacent fields, from AlphaFold leveraging the Protein Data Bank [4] to organic synthesis planning enabled by reaction databases [5], demonstrate that structured data repositories are prerequisites for predictive models. Yet inorganic materials science lacks comparable infrastructure. While recent efforts have utilized Large Language Models (LLMs) to extract synthesis text [6, 7], these datasets typically lack performance data, which is almost exclusively presented in charts and figures [8]. Without linking synthesis protocols to experimental outcomes, models cannot learn to optimize recipes for specific properties.

To bridge this gap, we present a multi-modal "Map of Materials Science", an extensible framework designed to curate linked synthesis-performance databases from open-access literature. Building on the `LeMat-Synth` toolbox [9], we introduce an advanced vision pipeline that leverages computer vision models and Vision-Language Models (VLMs) to digitize complex performance figures. We validate this approach by curating synthesis procedures spanning 35 methods and 16 material classes from 81,000 open-access papers [10]. In a targeted case study on thermocatalysis, our framework automatically extracts catalyst synthesis procedures and links them to their corresponding conversion and Arrhenius plots, establishing the necessary data infrastructure for autonomous laboratories and inverse materials design.

2. The Multi-Modal Extraction Framework

Our pipeline consists of two parallel extraction streams that converge to link procedural text with visual data, as shown in Figure 2.

2.1 Structured Synthesis Extraction

To extract synthesis protocols, we define a formal ontology that represents each procedure as a sequence of discrete operations [9]. We utilize the DSPy framework [11] to optimize an extraction pipeline using Gemini 2.0 Flash, which identifies target materials, parses multi-step procedures (including precursors, equipment, and conditions), and formats the output into a standardized JSON schema. This ontology is fully customizable, allowing the toolbox to adapt to specialized domains such as solid-state electrolytes or catalytic nanomaterials [12].

2.2 Performance Figure Digitization

Recognizing that performance data is predominantly visual, we developed a three-stage figure processing workflow. First, we employ DINO [13], a zero-shot visual segmentation model, to decompose multi-

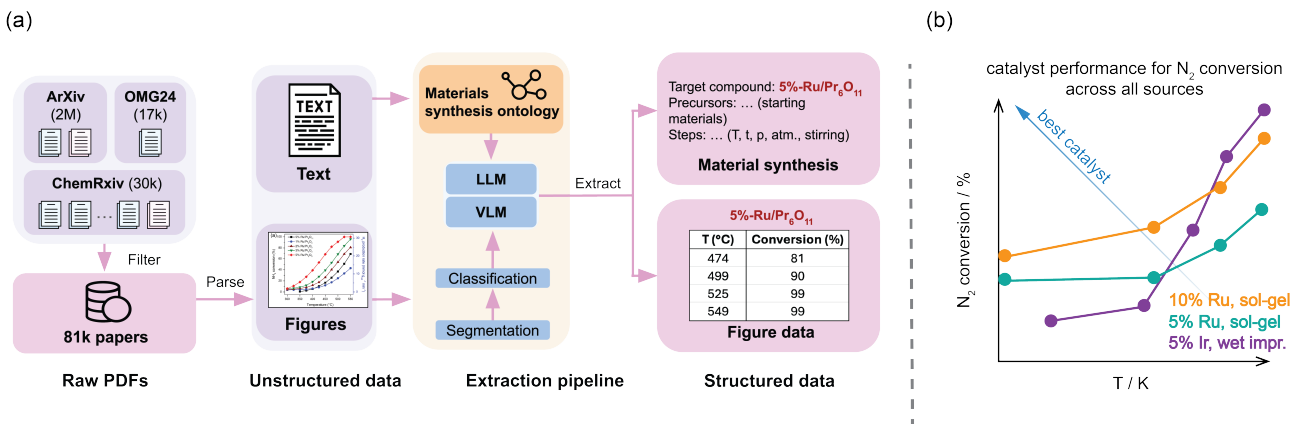

Fig. 2: **(a)** Overview of the `LeMat-Synth` pipeline. The framework processes unstructured PDFs through parallel streams: an LLM extracts structured synthesis ontologies (precursors, steps) while a VLM pipeline segments and digitizes performance figures. **(b)** The Map of Thermocatalysis. An aggregated visualization of digitized light-off curves extracted from the literature. The X-axis represents reaction temperature and the Y-axis represents conversion efficiency. Lines are colored by synthesis method and materials (e.g., sol-gel, wet impregnation), which reveals trends, performance clusters and identifying "best-in-class" synthesis strategies.

panel figures into individual subplots. Second, a fine-tuned ResNet-152 classifier filters these subplots to identify quantitative data charts (e.g., line plots, scatter plots) while discarding qualitative imagery like SEM micrographs [14]. Finally, we leverage Vision-Language Models (VLMs) such as Claude 3.5 Sonnet to digitize the identified charts, extracting numerical (x, y) coordinate pairs, axis labels, and units. This pipeline transforms raster images into machine-readable performance curves.

3. Case Study: Mapping Thermocatalysis

To demonstrate the utility of this multi-modal toolbox, we applied it to the domain of thermocatalysis, where linking catalyst preparation to activity is critical. We curated a focused subset of high-relevance publications from our broader corpus of 81,000 papers. The pipeline first classified papers to identify those containing both explicit catalyst synthesis recipes and catalytic performance metrics (e.g., conversion efficiency vs. temperature).

For each identified paper, the text pipeline identified the target output material for the synthesis and extracted the detailed procedure, specifying precursors, calcination temperatures, and support materials. Simultaneously, the vision pipeline targeted and digitized Arrhenius and light-off curves. A heuristic matching module then linked the extracted target material names (e.g., "5%-Ru/Pr_2O_3") with the corresponding legend labels in the digitized figures [15].Validation against a set of expert-annotated ground truth papers indicates that this automated "mapping" significantly reduces the manual effort required to curate property-performance datasets. By effectively coupling the synthesis graph with the performance graph, we enable the training of models that can predict not just the structure of a material, but its functional efficiency based on its manufacturing history.

Outlook

We present an extensible, open-source infrastructure for transforming unstructured scientific literature into machine-readable knowledge. By combining scalable synthesis extraction with automated figure digitization, we provide the community with the tools to build the "Map of Materials Science"—a connected database where every material is linked to both its creation and its capability. This work lays the foundation for predictive synthesis planning and supports the data needs of self-driving laboratories [16]. We release the `LeMat-Synth` toolbox[9], the curated dataset, and trained models as open-source infrastructure to enable the materials science community to build predictive synthesis models at scale.

Acknowledgments

We would like to acknowledge the volunteers of the LeMaterial community involved in this project: Xiyao Li, Ayush Jain, Georgia Channing, Gregoire Germain, Faris Flaifil, Alexandre Nozadze, Stefan Schmid, Mohd Zaki, Sudheesh Kumar Ethirajan, Elton Pan, N M Anoop Krishnan, Defne Circi, Haiwen Dai, Kedar Hippalgaonkar, Luis Pinto, Samuel Blouir, Thorben Prein, Vasily Safronov, Ilya Mazalov.

References

[1] Manoj Ravi and Joshua W Makepeace. Facilitating green ammonia manufacture under milder conditions: what do heterogeneous catalyst formulations have to offer? *Chem. Sci.*, 13(4):890–908, January 2022.

[2] Qifeng Zhang, Evan Uchaker, Stephanie L Candelaria, and Guozhong Cao. Nanomaterials for energy conversion and storage. *Chemical Society Reviews*, 42(7):3127–3171, 2013.

[3] Chang Liu, Feng Li, Lai-Peng Ma, and Hui-Ming

Cheng. Advanced materials for energy storage. *Advanced materials*, 22(8):E28–E62, 2010.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[5] Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

[6] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mat.*, 29(21):9436–9444, 2017.

[7] Zheren Wang, Kevin Cruse, Yuxing Fei, Ann Chia, Yan Zeng, Haoyan Huo, Tanjin He, Bowen Deng, Olga Kononova, and Gerbrand Ceder. Ulsa: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery*, 1(3):313–324, 2022.

[8] Maciej P. Polak and Dane Morgan. Leveraging Vision Capabilities of Multimodal LLMs for Automated Data Extraction from Plots.

[9] Magdalena Lederbauer, Siddharth Betala, Xiyao Li, Ayush Jain, Amine Sehaba, Georgia Channing, Grégoire Germain, Anamaria Leonescu, Faris Flaifil, Alfonso Amayuelas, Alexandre Nozadze, Stefan P Schmid, Mohd Zaki, Sudheesh Kumar Ethirajan, Elton Pan, Mathilde Franckel, Alexandre Duval, N M Anoop Krishnan, and Samuel P Gleason. LeMat-Synth: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature. October 2025.

[10] Heegyu Kim, Taeyang Jeon, Seungtaek Choi, Ji Hoon Hong, Dong Won Jeon, Ga-Yeon Baek, Gyeong-Won Kwak, Dong-Hee Lee, Jisu Bae, Chihoon Lee, Yunseo Kim, Seon-Jin Choi, Jin-Seong Park, Sung Beom Cho, and Hyunsouk Cho. Towards Fully-Automated Materials Discovery via Large-Scale Synthesis Dataset and Expert-Level LLM-as-a-Judge.

[11] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP.

[12] Elton Pan, Soonhyoung Kwon, Zach Jensen, Mingrou Xie, Rafael Gómez-Bombarelli, Manuel Moliner, Yuriy Román-Leshkov, and Elsa Olivetti. Zeosyn: A comprehensive zeolite synthesis dataset enabling machine-learning rationalization of hydrothermal parameters. *ACS Cent. Sci.*, 10(3):729–743, 2024.

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[14] KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.

[15] Diego Mateo, Angel Sousa, Maksim Zakharzhevskii, and Jorge Gascon. Challenges and opportunities for the photo-(thermal) synthesis of ammonia. 26(3):1041–1061.

[16] Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. Self-driving laboratories for chemistry and materials science. *Chem. Rev.*, 124(16):9633–9732, August 2024.