ChemComm

HIGHLIGHT

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2025, **61**, 11083

Text mining in MOF research: from manual curation to large language model-based automation

Suyeon Bae,[a] Mingyu Jeon *[b] and Hoi Ri Moon *[a]

The rapid expansion of metal–organic framework (MOF) literature presents both a rich resource and a significant challenge for knowledge extraction. Text mining, which enables the conversion of unstructured scientific texts into structured, machine-readable data, has emerged as a key tool for accelerating data-driven research in the MOF domain. This review traces the development of text mining approaches in MOF research, from early manual curation and rule-based methods to recent breakthroughs powered by large language model (LLM)-based automation. We discuss the foundational role of natural language processing (NLP) and machine learning (ML) techniques such as named entity recognition and vector embedding models, followed by an in-depth analysis of LLM-based frameworks that enable flexible, scalable, and context-aware information extraction. Additionally, we introduce and compare their accuracy, and explore their diverse applications—including prediction of synthesizability, materials properties, and thermal stability. We conclude with a perspective on future directions for text mining in MOF research, including its integration into interactive graphical user interfaces, autonomous laboratories, multi-agent AI systems, and multi-modal LLM frameworks that can process textual, visual, and structural information in a unified way. This review aims to provide a foundational understanding for both experimental and computational researchers interested in adopting or advancing text mining methods in the MOF field.

Received 2nd May 2025,
Accepted 26th June 2025

DOI: 10.1039/d5cc02511g

rsc.li/chemcomm

1. Introduction

Metal–organic frameworks (MOFs) represent one of the most versatile and innovative material classes developed in recent

[a] *Department of Chemistry and Nanoscience, Ewha Womans University, Seoul, 03760, Republic of Korea. E-mail: hoirimoon@ewha.ac.kr*
[b] *Computational Science Research Center, Korea Institute of Science and Technology, Seoul, 02792, Republic of Korea. E-mail: mingyu1116@kist.re.kr*

Suyeon Bae

Suyeon Bae received her BS degree from Ewha Womans University in 2025. She is currently pursuing her MS degree at the NanoBio Electrochemistry Laboratory at Ewha Womans University. Her research interests include the development of next-generation energy materials based on nanocatalysts and electrochemical analysis.

Mingyu Jeon

Mingyu Jeon received his PhD in Chemical and Biomolecular Engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2024. He is currently a postdoctoral researcher at the Korea Institute of Science and Technology (KIST). His doctoral research focused on computational modeling and discovery of π-conjugated 2D conductive MOFs towards chemiresistive gas sensor applications. His current interests include developing AI-driven tools for designing superionic solid-state electrolytes, with machine learning potentials and generative models.

decades. Formed through the coordination of metal ions or clusters with organic ligands, MOFs exhibit an exceptional combination of high porosity, tunable chemical functionality, and structural adaptability.[1,2] These distinctive characteristics have propelled MOFs to the forefront of materials science, positioning them as key enablers in addressing critical challenges in energy, environmental sustainability, and biomedical applications.[2–6] Furthermore, their highly customizable and tunable nature allows for precise structural and chemical modifications, reinforcing their significance in both fundamental research and industrial application.

Despite their immense potential, the structural diversity that makes MOFs highly attractive also introduces significant challenges. Given the vast number of synthesized MOFs and computationally generated hypothetical MOFs, exploring optimal materials for specific applications has become increasingly complex.[7,8] This challenge is further compounded by the inherent limitations of conventional trial-and-error approaches and the labor-intensive nature of experimental validation. In response, researchers have increasingly advocated for systematic, data-driven methodologies to effectively navigate the vast chemical landscape of MOFs and accelerate the discovery of application-specific materials (Scheme 1).[9,10]

To overcome these challenges, text mining has emerged as a powerful approach for systematically analyzing large-scale scientific literature. By applying natural language processing (NLP) techniques, researchers can extract structured and informative data from unstructured text, tables, and figures.[11–13] The systematic compilation of information on MOF synthesis conditions, experimental methodologies, and performance metrics facilitates the construction of high-quality databases, establishing a robust foundation for the accelerated discovery of novel materials. Beyond streamlining data extraction and enabling automated database curation, text mining also aids in identifying emerging research trends and uncovering previously overlooked structure–property relationships within the literature.[9,14]

Hoi Ri Moon

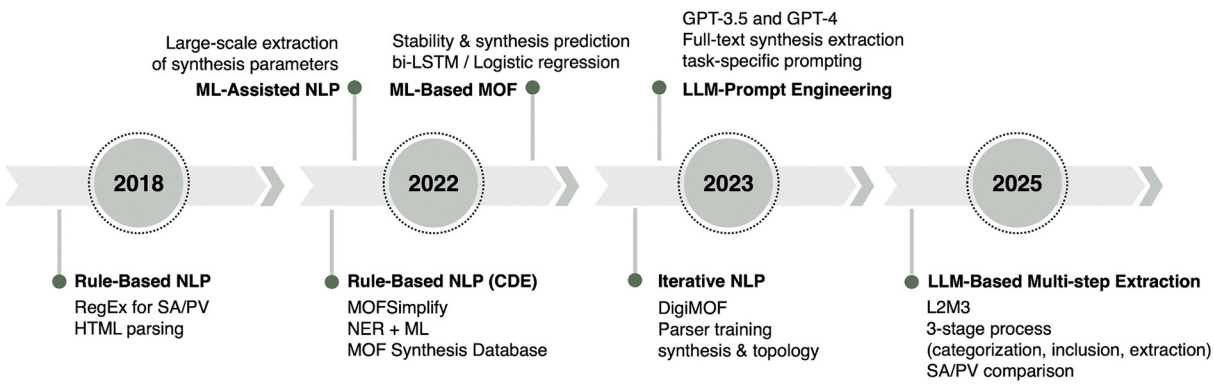
Hoi Ri Moon is a professor in the Department of Chemistry and Nanoscience department at Ewha Womans University joined in 2023 as a Ewha Fellow professor. In 2007, she completed her PhD degree in Chemistry at Seoul National University. Following that, she conducted postdoctoral research at Molecular Foundry of Lawrence Berkeley National Laboratory. From 2010 to 2022, she served as a faculty member at UNIST. She is a vice chair of Metal–Organic Frameworks-International Commission. Her current research focuses on ordered–disordered and flexible MOFs for applications such as gas separation, storage, molecular sensing, and catalytic reactions.

The most straightforward and fundamental text mining approach is manual curation by researchers, which involves searching for relevant publications, identifying those that align with a specific research focus, and extracting key textual elements such as paragraphs, sentences, and relevant terms. However, this method heavily depends on domain expertise, making it less scalable and efficient for widespread use. Moreover, with the rapid expansion of published literature, manually processing such a vast amount of information is increasingly impractical, highlighting the necessity for more automated and systematic approaches.[15–17]

The breakthroughs of NLP and machine learning (ML) methodologies has transformed text mining research by enabling enhanced automation and more precise data extraction. Techniques such as Word2Vec, Paragraph2Vec, and Paper2Vec facilitate the automated selection and classification of research papers.[18–20] Named entity recognition (NER) using bidirectional long short-term memory (Bi-LSTM) ML techniques improve the classification, identification, and extraction of key information from unstructured text based on user-defined labels, further improving data accessibility and usability.[21–23] With the emergence of the transformer architecture in 2017,[24] text mining research underwent a significant advancement, driving the development of transformer-based models such as bidirectional encoder representations from transformers (BERT).[25] BERT has since been adapted in various chemistry and materials science domains through specialized models like MatBERT,[26] SciBERT[27] and BatteryBERT[28] substantially enhancing automation, efficiency, and accuracy compared to earlier Bi-LSTM-based NER models. Despite their advantages, rule-based methods still required manual curation, limiting them to partially automated workflows and single-purpose tools designed for domain experts. These approaches struggled to handle the complexity and diversity of scientific literature.

The advent of large language models (LLMs), pretrained on vast datasets, has driven the innovation in text mining research.[29–32] LLMs such as GPT-3.5, GPT-4[33] Gemini1.5,[34] and Llama3.1[35] demonstrate the ability to tackle tasks in chemistry and materials science, even without explicit domain-specific training. Integrating LLMs into text mining facilitates a more comprehensive and automated data extraction process while offering users with greater flexibility in decision-making. Recent studies have explored fine-tuning of LLMs with prompt engineering using small, domain-specific chemical knowledge datasets—consisting of only a few dozen samples—to further enhance the performance and adaptability of LLMs.[36–38] A significant development in this area has been the emergence of iterative NLP workflows, where LLM-based models undergo repeated cycles of extraction, error correction, and rule refinement to enhance precision and recall in multi-step information harvesting.

In this feature article, we introduce the role of text mining in MOF research, with a particular focus on data extraction techniques and their impact on scientific discovery. We begin with rule-based text mining, which relies on human intervention through conventional NLP and ML approaches to extract



Scheme 1 Timeline showing the evolution of text mining in MOF research, from rule-based NLP (2018) to ML-based approaches (2022), and LLM integration (2023–2025). Ref. 23, 40, 43, 46, 47, 56 and 57.

relevant information. We then review the latest advancements in LLM-based text mining, highlighting how LLMs have transformed methodologies and research trends of rule-based text mining. Finally, we discuss key insights and future directions for integrating text mining into MOF research. Our aim is to make text mining more accessible to both experimental and computational MOF researchers, facilitating its seamless adoption into their workflows and accelerating data-driven discoveries.

2. Machine learning and rule-based text mining in MOF research

2.1 Performance and stability descriptor extraction

Before the advent of rule-based or machine learning-driven text mining techniques, early efforts in MOF data extraction relied almost exclusively on manual curation by domain experts.[39] This foundational approach involved researchers meticulously identifying relevant publications, carefully examining experimental sections, and extracting key textual information, numerical values, and structural descriptors into organized formats, often spreadsheet-like databases. While inherently labor-intensive and limited in scalability—heavily depending on the specialized knowledge of individual researchers—this method was critical for establishing the initial landscape of MOF research.

Crucially, this considerable manual input provided the high-quality, reliable foundational data that underpinned the development of early MOF databases. These manually assembled datasets, such as early iterations of the computation-ready, experimental (CoRE) MOF Database and meticulously curated subsets within the broader Cambridge Structural Database (CSD), served as invaluable ground truth for validating subsequent automated text mining and data extraction systems (Table 1). The meticulous human oversight ensured the fidelity and chemical correctness of the extracted information, which was paramount for the nascent stages of computational MOF research and played a pivotal role in developing this area of research. This legacy of expert-driven manual efforts continues to inform current practices, with hybrid approaches combining automated logic with expert oversight remaining vital in today's semi-automated data pipelines, particularly for validation and handling complex cases.

Early NLP methodologies, predominantly rule-based approaches, relied on pre-defined heuristics and keyword-based extraction techniques. While effective in well-structured text formats, these methods struggled with linguistic variability and the complexity of scientific discourse. To overcome these limitations, such ML techniques have been incorporated into NLP workflows, enabling more flexible and scalable data extraction.

In 2018, the earliest application of text mining to MOFs was conducted by Kim *et al.*, who developed a rule-based extraction system using regular expressions (RegEx) to retrieve surface area (SA) and pore volume (PV) from MOF-related literature.[40] This algorithm was specifically designed to work with articles as hypertext markup language (HTML) format and employed RegEx to detect numerical values associated with SA and PV by identifying their commonly used units (*e.g.*, $m^2\ g^{-1}$ for SA and $cm^3\ g^{-1}$ for PV).

The study's workflow consisted of HTML parsing, text tokenization, keyword filtering, and unit detection. Beautiful Soup 4.0 python library was used to preprocess HTML documents, eliminating irrelevant tags and extracting meaningful text. The algorithm then categorized tokens into four groups—MOF name, unit, numerical value, and keyword—to systematically match SA and PV data to the correct MOF structures (Fig. 1a). A key challenge addressed in this approach was that MOF names were

Table 1 Overview of major metal–organic framework databases. Contents and access information for key MOF repositories, indicating CIF availability, included properties, and URL or DOI reference

Database name	CIF included?	Additional properties	Access
CoRE MOF	Yes	Experimental SA, PV, density	https://doi.org/10.5281/zenodo.3677685
CSD MOF subset	Yes	Crystallographic metadata, topology	https://www.ccdc.cam.ac.uk/(CSD access)

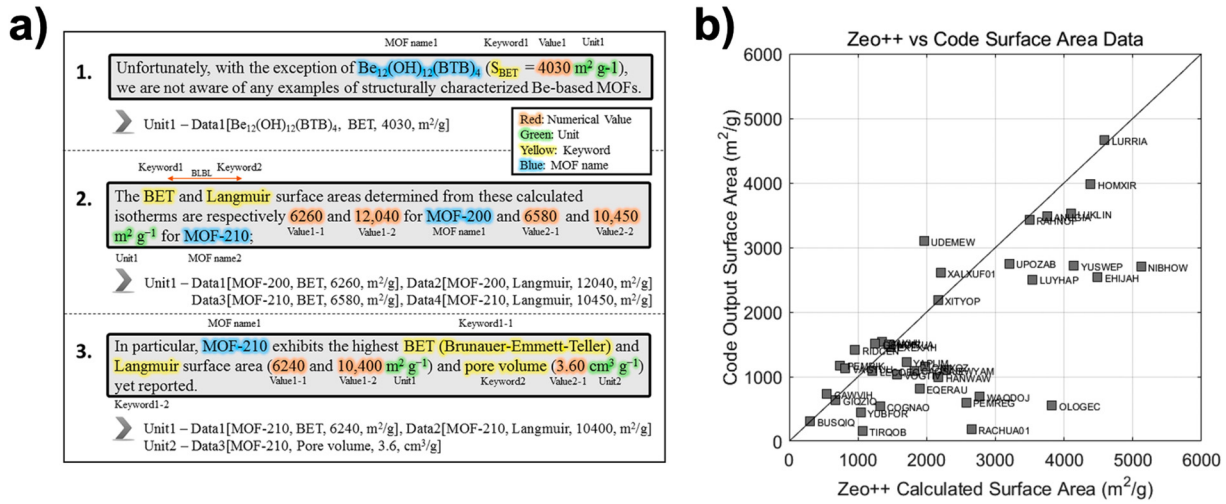


Fig. 1 (a) Example cases illustrating the identification scheme of a rule-based text mining algorithm. The algorithm extracts surface area (SA) and pore volume (PV) values from scientific literature by recognizing MOF names, numerical values, keywords, and units, distinguishing between BET and Langmuir SA types, and correctly mapping data points. (b) Comparison between surface area values extracted by the algorithm (code output) and those calculated using Zeo++ for various MOFs. A strong correlation indicates the reliability of the extraction process. Reprinted from ref. 40 with permission from American Chemical Society, Copyright 2018.

often not located in the same sentence as their corresponding SA or PV values. To address this challenge, the algorithm searched for up to four sentences forward and backward, ensuring a more accurate data mapping.

To further assess its accuracy, the algorithm was applied to 2315 HTML files from the CoRE MOF database,[41,42] an extensive collection of experimentally synthesized MOFs. This database, largely derived from the CSD, shows how extensively MOF collections have been curated and organized for computational applications. The CSD, a widely popular repository of crystallographic information, has drawn considerable attention since incorporating a comprehensive MOF subset. Notably, this subset is recognized as the world's first automatically updated MOF dataset, containing nearly 100 000 structures as of 2020.[43] The development of the CSD MOF subset marks a significant advancement in automated, large-scale data extraction and database maintenance. This infrastructure enables targeted searches and detailed analysis of various MOF properties.

From the complete set of articles in HTML format, the system extracted 490 SA values and 250 PV values, establishing a comprehensive dataset for computational modelling. The remaining documents either did not report SA/PV values explicitly (for example, presenting only graphical adsorption isotherms without accompanying numerical data), employed non-standard units or abbreviations beyond the predefined regex patterns, or exhibited HTML structures that impeded reliable parsing.

Given the impracticality of manually validating each extraction, we randomly selected 50 papers for spot-checking. In these cases, 134 of the 183 SA values (73.2%) and 200 of the 235 PV values (85.1%) matched the manually curated reference, compared to the 90% (SA) and 88.8% (PV) accuracies previously obtained on curated review papers.

Additionally, the text-mined surface areas were compared with Zeo++-calculated values for the same collection of structures (Fig. 1b). While a general linear trend is evident, several outliers in the bottom-right quadrant indicate instances where the simulation overestimated experimental BET results. These deviations likely stem from various factors, including incomplete solvent removal in laboratory measurements, framework distortions during activation, or the idealized pore geometries assumed by Zeo++.[44]

While the observations presented here highlight areas for improvement in future work, particularly regarding the handling of unit formats and structural complexities, the overall consistency with simulation data demonstrates the utility of rule-based text mining for extracting MOF performance properties from unstructured literature. An important advance in this context has been the emergence of iterative NLP workflows, where LLM-based models undergo repeated cycles of extraction, error correction and rule refinement to improve both precision and recall in multi-step information harvesting.

To overcome the limitations of purely rule-based methods, subsequent work has focused on integrating machine learning models with NLP techniques to boost both extraction accuracy and predictive power. For example, Kulik *et al.* introduced MOFSimplify in 2022, combining ChemDataExtractor (CDE)[45] with ML classifiers to extract and predict stability-related descriptors.[43] This study utilized NER and dependency parsing to retrieve solvent removal stability and thermal decomposition temperatures from scientific texts (Fig. 2a). The resulting dataset encompassed 2179 solvent removal stability entries and 3132 thermal stability annotations, forming one of the largest MOF stability datasets curated through NLP.[43]

A significant advancement introduced in this study was the integration of user-contributed thermogravimetric analysis (TGA) traces, allowing direct validation of text-mined data against experimental results (Fig. 2b). Comparative analyses with manually curated datasets confirmed the high accuracy of

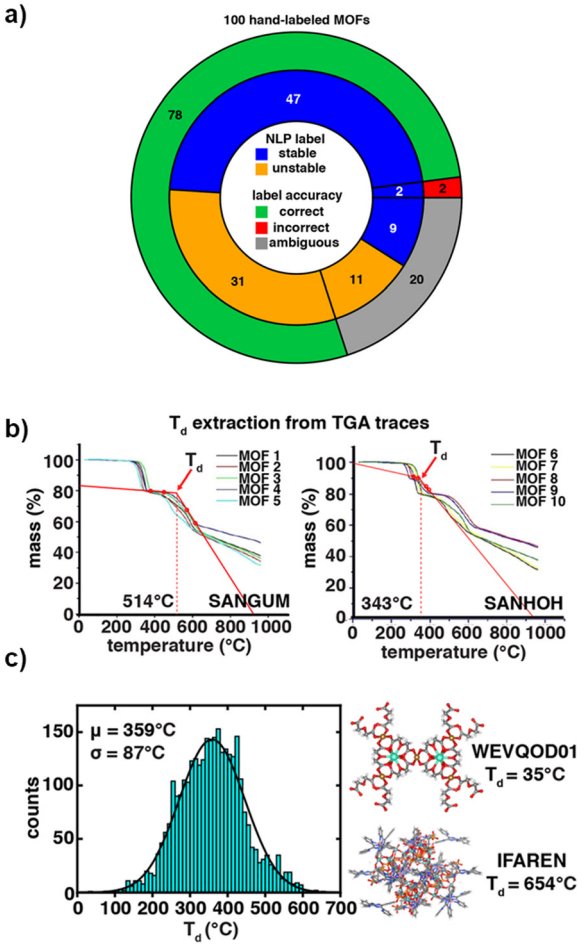


Fig. 2 Validation and application of machine learning (ML)-driven text mining for solvent removal stability and thermal decomposition temperature prediction. (a) Comparison of NLP-assigned stability labels to manually assigned labels for 100 MOFs, where correctly classified cases are marked in green, incorrect assignments in red, and ambiguous cases in gray. (b) Extraction of decomposition temperature (T_d) from thermogravimetric analysis (TGA) traces for selected MOFs (SANGUM and SANHOH), highlighting variations in thermal stability. (c) Distribution of extracted decomposition temperatures (T_d) for the full dataset, with representative MOFs exhibiting the lowest (WEVQOD01) and highest (IFAREN) thermal stability. Reprinted from ref. 43 with permission from Nature, Copyright 2022.

the NLP-assigned stability labels, while the extracted TGA-derived decomposition temperatures exhibited strong correlation with manually annotated values. For instance, MOFs having refcode of SANGUM and SANHOH in the CSD demonstrated decomposition temperatures of 514 °C and 343 °C, respectively, highlighting the capability of automated NLP approaches in retrieving experimental stability data from the literature (Fig. 2b). The distribution of thermogravimetric analysis TGA-derived decomposition temperatures for MOFs reveals a normal distribution centered around 359 °C with a standard deviation of 87 °C (Fig. 2c). This visualization highlights the variability in MOF thermal stability and validates the robustness of the NLP-extracted dataset through systematic temperature extraction.

To further utilize the extracted data, artificial neural networks (ANNs) were trained using the mined stability dataset, achieving over 90% accuracy in predicting solvent removal stability and decomposition temperatures. This study exemplifies how the synergy between text mining and ML enables the transformation of literature-derived MOF descriptors into predictive modelling frameworks. In addition to stability assessment, ML and NLP-driven text mining has facilitated the extraction of MOF synthesis conditions, including reaction temperatures, solvents, and metal precursors. Kim *et al.* pioneered an NLP-based system that extracted synthesis-relevant information from 28 565 MOF-related publications.[23] This study utilized logistic regression, support vector machines (SVM), and random forest models for synthesis paragraph classification, with logistic regression achieving the highest precision (>98%) in identifying synthesis-related passages. Within the synthesis paragraph, bi-LSTM combined with conditional random field (CRF) layer was used to extract and categorize the relevant chemicals. Using the extracted dataset, an ANN was trained with positive-unlabeled (PU) learning to assess whether specific synthesis conditions would enable successful synthesis. This text mining study enables researchers to facilitate ideal synthesis conditions and predict synthesizability based on literature patterns.

2.2 Synthesis condition extraction and database construction

The use of natural language processing (NLP) in MOF research has significantly improved the extraction and analysis of synthesis conditions. Earlier studies primarily focused on extracting performance-related descriptors, such as thermal stability[43] or identifying synthesis-relevant sections from scientific texts.[23] More recent efforts have shifted toward constructing large-scale databases that systematically organize MOF synthesis parameters. This transition is necessary to address the limitations of manually curated datasets, which often restrict the scalability of computational approaches for MOF synthesis analysis.

One of the most comprehensive implementations of this large-scale text mining approach is DigiMOF, which applies rule-based NLP parsing using ChemDataExtractor (CDE) to systematically structure MOF synthesis data.[46] To ensure extraction accuracy, DigiMOF employs an iterative parser training process, where text mining rules are refined and validated through continuous feedback (Fig. 3a). This iterative refinement allows the database to improve precision while integrating newly published MOF synthesis studies.

DigiMOF extracts key synthesis parameters, including solvents, metal precursors, and organic linkers, from a dataset of over 43 000 scientific publications. The database construction follows a structured pipeline designed for efficient and accurate data retrieval. Initially, digital object identifiers (DOIs) linked to MOF-related publications were automatically retrieved from the CSD MOF subset. The extracted documents then underwent preprocessing steps, including tokenization, part-of-speech (POS) tagging, and chemical entity recognition, to segment and classify relevant textual components. This process improves the accuracy of parameter identification and minimizes classification errors.

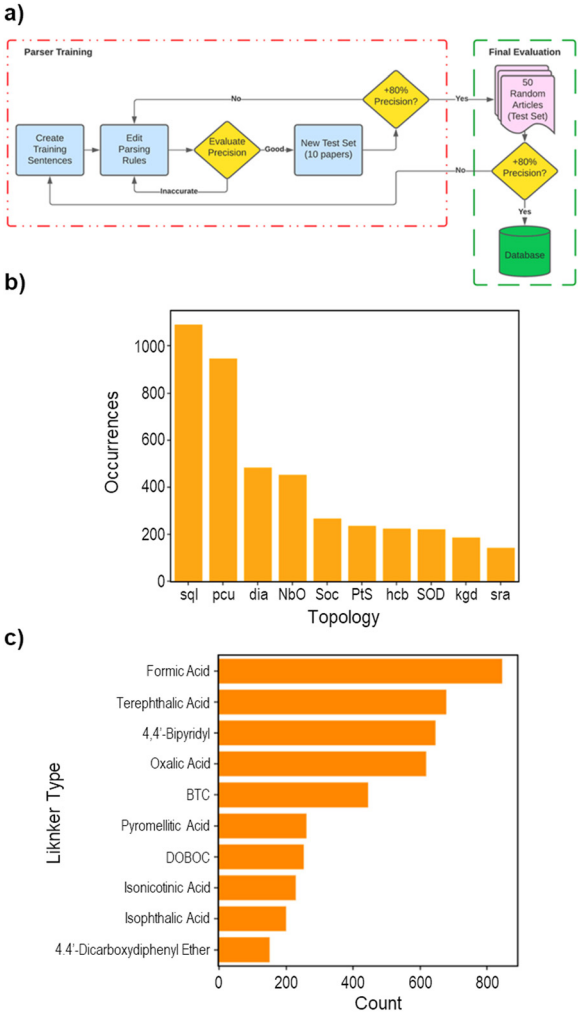


Fig. 3 Workflow and topological analysis of MOFs extracted through text mining. (a) Iterative parser training process, where extraction rules are refined and evaluated for precision until accuracy exceeds 80%. (b) Histogram of the most frequently occurring MOF topologies identified using ChemDataExtractor (CDE), with **sql** and **pcu** being the most common. (c) Histogram of the most frequently occurring MOF topologies extracted from 3D structures using CrystalNets. Reprinted from ref. 46 with permission from American Chemical Society, Copyright 2023.

To assess the reliability of the extracted data, a comparative analysis with manually curated datasets was conducted, confirming the high reliability of the NLP-based extraction process. This validation step ensured that the structured synthesis data in DigiMOF aligned well with known synthesis conditions. The final dataset contains 52 680 synthesis property relationships across 15 501 unique MOFs, covering approximately 15% of the CSD MOF subset. This automated text mining approach facilitates the generation of a high-quality database that integrates MOF synthesis data for future predictive modeling and high-throughput materials screening.

In addition to data extraction, DigiMOF's corpus quantifies topology usage at an unprecedented scale, confirming the dominance of **sql** and **pcu** frameworks while cataloguing 112 distinct topologies (Fig. 3b). The co-occurrence of synthesis parameters including solvent, temperature, and additive with topology and linker data enables multivariate correlation analyses that may uncover subtle protocol–structure relationships and inform targeted experimental design. Likewise, linker-occurrence mapping (Fig. 3c) verifies the predominance of carboxylate and pyridyl ligands and also reveals less common chemistries, such as azolate, warranting further investigation. By structuring these extensive datasets, DigiMOF establishes a foundation for data-driven hypothesis generation and the subsequent development of predictive machine-learning frameworks for MOF synthesis.

While DigiMOF provides a structured repository of MOF synthesis conditions, further efforts have focused on refining text mining techniques to extract synthesis-specific parameters with greater accuracy. Tsotsalas *et al.* developed such an approach, implementing a multi-step workflow to systematically extract MOF synthesis parameters from scientific literature.[47]

Tsotsalas *et al.* applied a structured text mining approach to systematically extract MOF synthesis parameters, beginning with the collection of 6099 journal articles from major publishers. First, a paragraph classification step was conducted using a decision tree-based string search method to automatically select synthesis-related sections from this large corpus, significantly reducing the need for manual curation and improving both efficiency and scalability. Next, the ChemicalTagger software was applied to the selected paragraphs to identify and extract key synthesis parameters, including solvents, reaction temperatures, additive use, and reaction times.

After identifying relevant text, ChemicalTagger, an NLP tool designed for parsing experimental procedures, was used to extract key synthesis parameters, including solvent, reaction temperature, additive use, and reaction time. To improve accuracy, domain-specific modifications were made to the NLP pipeline, ensuring proper recognition and classification of MOF-related terminology, such as coordination environments, solvent polarity effects, and metal precursor names. Additionally, crystallographic information files (CIFs) were obtained from two well-curated structural repositories—the CoRE MOF and the CSD—and analyzed to extract structural attributes such as metal-center oxidation states, linker compositions, and framework connectivity (Fig. 4a).

To validate the accuracy of the extracted data, a comparative analysis with manually curated datasets was performed. The dataset was further analyzed to identify trends in synthesis parameter relationships. The temperature–solvent–additive relationships with DMF and water dominating the 80–160 °C range (Fig. 4b), water being universally used above 160 °C (consistent with hydrothermal methods), and acidic additives largely limited to syntheses below 80 °C, are well established in MOF chemistry. However, automated text-mining at scale quantifies how frequently each protocol occurs across more than 6000 publications and reveals unusual instances, such as high-temperature syntheses using acidic additives, that deviate from conventional practice. Furthermore, these comprehensive statistics serve as a resource for generating hypotheses, drawing attention to underexplored solvent–additive combinations.

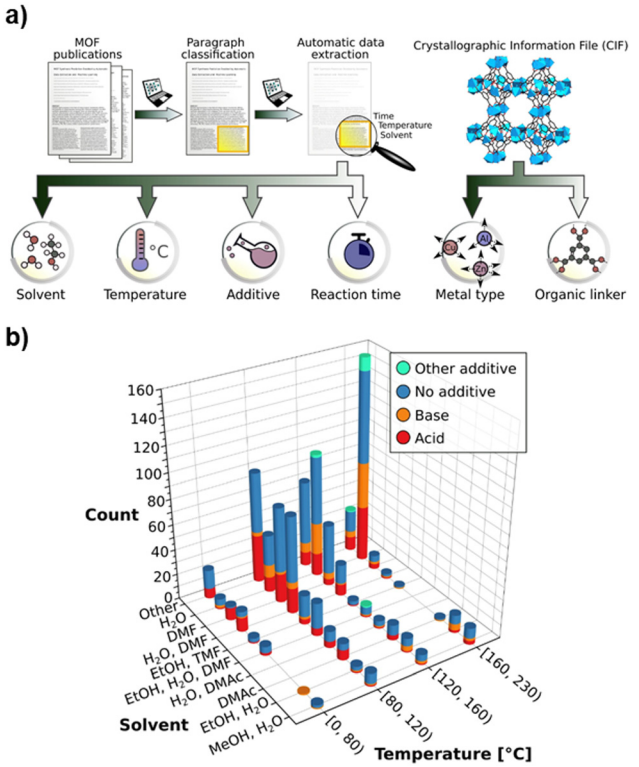


Fig. 4 (a) Text-mining pipeline for extracting MOF synthesis parameters from literature. Synthesis-relevant paragraphs are first identified, then tagged using ChemicalTagger to extract parameters including metal source, linker, solvent, additive, temperature, and synthesis time. (b) Statistics of the SynMOF database constructed from the extracted information: frequency of different metal sources, most commonly used linkers, and their structural diversity. Reprinted from ref. 47 with permission from Wiley, Copyright 2022.

Importantly, incorporating this structured dataset into machine-learning frameworks such as SynMOF enables predictive synthesis capabilities, thereby accelerating the discovery and optimization of novel MOF synthesis routes.

Using the structured dataset obtained through text mining, the study further explored the application of machine learning (ML) models to predict MOF synthesis conditions, including reaction temperature and solvent selection. The SynMOF database, established through this automated text mining approach, served as the basis for training ML models for synthesis condition prediction. However, the primary focus remained on refining text mining techniques to achieve accurate extraction of synthesis parameters. This work demonstrates the potential of combining text-mined synthesis data with computational models to assist in guiding MOF synthesis strategies.

3. LLM-based text mining

Despite significant improvements in accuracy and adaptability of text mining achieved through ML and NLP techniques, rule-based approaches still required extensive domain-specific feature engineering and struggled to handle the variability and complexity of scientific language. To address these limitations, recent advancements in LLMs have introduced a more flexible and context-aware approach to text mining. Unlike rule-based text mining approaches, which required extensive human intervention, LLMs have transformed knowledge extraction by enabling interpretation of complex information without requiring extensive manual rules. They can process and analyse text with minimal examples (zero-shot or few-shot learning) or be fine-tuned for specialized domains, allowing for flexible and adaptive data understanding.

The application of LLMs is expanding rapidly in various materials systems, beyond MOF research. Very recently, Lee *et al.* introduced a language modeling-based protocol, text-to-battery recipe (T2BR), for the automated extraction of complete battery material recipes—from synthesis to cell assembly—by integrating ML-based NLP and LLMs.[48] Through the construction of a structured dataset comprising 165 end-to-end recipes, the study enabled the identification of trends such as precursor–method associations. In the field of water-splitting catalysis, Kim *et al.* developed MaTableGPT, an LLM-based framework for extracting complex and diverse tabular data from scientific literature.[13] By introducing two key strategies—table data representation and table splitting—they improved GPT comprehension and effectively filtered hallucinated information. Notably, the few-shot learning approach emerged as the most balanced solution, offering both a high extraction score (nearly 95% total F1 score) and low cost (GPT usage cost of 5.97 US dollars and labeling cost of 10 I/O paired examples). Furthermore, Jain *et al.* developed a LLM-based framework for extracting structured scientific knowledge from text, with a focus on diverse materials domains: dopant–host relations, MOFs, and general composition/phase/morphology/application relationships.[49] By fine-tuning pre-trained LLMs—OpenAI's GPT-3 (closed source) and Meta's LLama-2 (open source)—they achieved high performance in joint NER and relation extraction (NERER), accurately transforming complex and hierarchical information into structured formats like JSON.

While LLMs have shown great promise in academic research for extracting structured scientific knowledge, their impact is also extending rapidly into the industrial sector. Very recently, Sattar *et al.* provide a comprehensive overview of how LLMs are transforming industry by automating complex natural language tasks, delivering high accuracy in data mining, and decision-making.[50] Applied across sectors such as medical,[51] automotive,[52] education,[53] e-commerce,[54] and finance,[55] LMs enable applications ranging from predictive diagnostics and fraud detection to personalized learning and real-time language translation. Aforementioned studies collectively underscore the pivotal role of LLMs in both various material science and industrial domains, highlighting a potential to further integrate LLMs within MOF science. LLM-driven text mining into MOF fields can facilitate the extraction of synthesis conditions, prediction of material properties, and large-scale dataset generation.

In 2023, Yaghi *et al.* introduced a ChatGPT-based LLM framework (GPT-3.5 and GPT-4) specifically designed for text mining in MOF chemistry, with a primary focus on extracting synthesis parameters from MOF-related publications.[56] By

using prompt engineering with chemistry-related tasks, researchers developed a ChatGPT chemistry assistant (CCA). To construct CCA, the study introduced a systematic prompt engineering approach, termed ChemPrompt Engineering, which was central to enabling domain-specific information extraction in a controlled and reproducible manner. The framework consists of three core steps: (1) minimizing hallucination by designing role-based prompts that clearly define ChatGPT's task and scope as a chemistry assistant; (2) providing task-specific instructions that guide the model to extract only relevant synthesis parameters—such as metal sources, linkers, solvents, temperatures, and reaction times—from varied experimental contexts; and (3) structuring the output format to ensure consistency and usability, typically in tabulated or JSON-style entries. This strategy not only improved the accuracy and interpretability of extracted information but also demonstrated that LLMs, when guided by domain-adapted prompts, can serve as scalable alternatives to traditional rule-based text mining systems in chemical literature analysis.

This model processes full-text research articles, automatically identifies key synthesis parameters such as metal sources, linkers, solvents, reaction temperature, and reaction time. In its initial validation, CCA was applied to a curated corpus of 228 MOF research articles (and their 225 supporting documents), yielding 2387 unique synthesis condition relationships. On this set, CCA achieved true positive counts exceeding 2000 for most parameter categories, demonstrating high extraction precision across metal source, linker, solvent, reaction temperature, and reaction time (Fig. 5a).

Furthermore, performance evaluations across three independent extraction processes revealed consistently high precision, recall, and F1 scores, highlighting the robustness of LLM-based text mining approaches in handling complex scientific language (Fig. 5b). In this study, the three processes—process 1 (sentence-level extraction), process 2 (paragraph-level summarization), and process 3 (multi-step extraction combining classification, summarization, and structuring)—were designed to test the model's adaptability to different input formats and task complexities. The consistently high performance across all three processes underscores the flexibility of CCA in processing scientific texts under varying levels of context and abstraction.

To demonstrate scalability, the pipeline was subsequently deployed across approximately 800 unique MOF structures, extracting 26 257 distinct synthesis parameter instances from peer-reviewed publications. Compared to conventional rule-based data mining methods, the CCA has demonstrated the potential for a more flexible and scalable approach to processing unstructured synthesis descriptions. LLM-based text extraction enables the creation of large-scale MOF synthesis databases, facilitating data-driven materials discovery and predictive synthesis modelling.

The ability to process vast amounts of scientific literature is a key advantage of LLMs over traditional NLP and ML-based text mining techniques. One of the most significant demonstrations of this capability is the very recent study by Kim *et al.*,

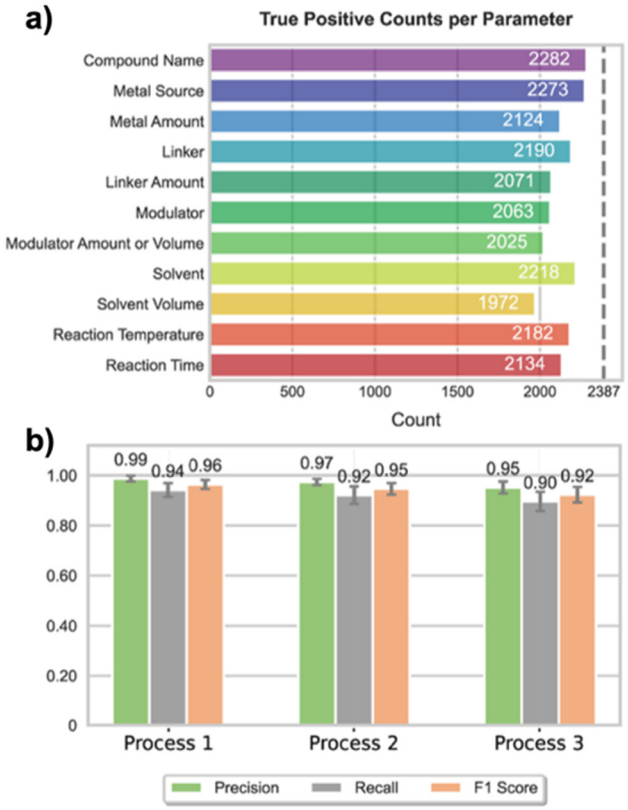


Fig. 5 Performance evaluation of text-mining processes for extracting MOF synthesis parameters. (a) True positive counts for 11 synthesis parameters, including compound name, metal source, linker, solvent, reaction temperature, and reaction time, demonstrating the accuracy of parameter extraction across 2387 synthesis conditions. (b) Comparison of precision, recall, and F1 scores across three different text-mining processes, showing consistently high performance with minor variations. Standard deviations are represented by gray error bars. Reprinted from ref. 56 with permission from American Chemical Society, Copyright 2023.

which implemented a LLM framework to extract and categorize MOF synthesis data from 41 681 scientific papers.[57]

To handle this large-scale dataset, the study employed a systematic pipeline consisting of three core tasks: categorization, inclusion, and information extraction. First, the model classified whether each paragraph was relevant to MOF synthesis (categorization task), followed by a decision on whether the synthesis information in the paragraph was complete enough to include in the dataset (inclusion task). Finally, for the paragraphs that passed both stages, detailed synthesis parameters such as metal sources, organic linkers, solvents, and additives were extracted using structured prompts (extraction task). The LLM achieved high F1 scores across all three tasks, with especially strong performance in the categorization and extraction steps, demonstrating the model's ability to process highly unstructured experimental text with minimal rule-based intervention.

The resulting dataset, compiled from synthesis-relevant paragraphs, encompasses detailed information on synthesis conditions and material properties. Statistical analysis of the mined data revealed meaningful trends: solvent types were the

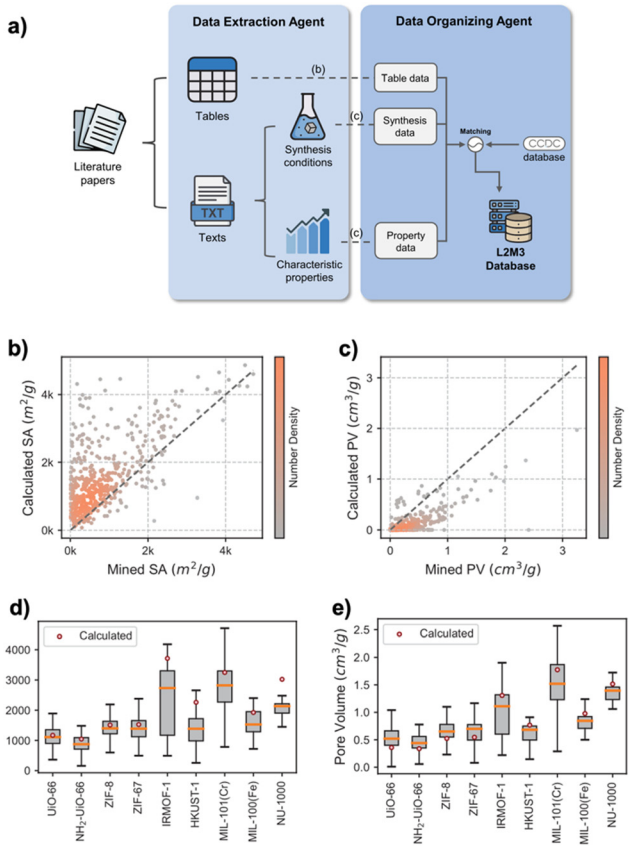


Fig. 6 (a) Overview of the L2M3 data extraction and organization workflow. Literature papers are processed by a data extraction agent that identifies and extracts information from tables and text, including synthesis conditions and characteristic properties. Extracted data are then structured and matched with entries in the CCDC database by the data organizing agent to build the L2M3 database. (b) Scatter plot comparing surface area (SA) values extracted by L2M3 with those calculated from MOF crystal structures. (c) Scatter plot comparing text-mined and calculated pore volume (PV) values. Color gradient represents the number density of data points. (d) Box plot of mined SA values for nine representative MOFs; red dots indicate the corresponding calculated values. (e) Box plot of mined PV values for the same MOFs, highlighting distribution and deviation from simulation-derived values. Reprinted from ref. 57 with permission from American Chemical Society, Copyright 2025.

most frequently extracted, followed by metal sources, linkers, and additives. Furthermore, compound-wise statistics showed that a large portion of MOFs were associated with multiple synthesis records, reflecting the diversity of experimental conditions under which the same material can be synthesized. The authors also analysed the distribution of synthesis data by publication year and journal, highlighting the steady increase in MOF synthesis reports and the broad coverage of the mined dataset across the chemical literature.

The use of LLM-based text mining enables the construction of large-scale experimental property datasets that were previously difficult to compile using manual or rule-based methods. Leveraging this capability, the authors performed a large-scale comparison between text-mined experimental values and simulation-derived values for surface area (SA) and pore volume (PV), allowing for a more systematic evaluation of their consistency (Fig. 6b and c). The analysis revealed notable discrepancies between the two data sources. While simulation values were consistent and singular for each MOF structure, the experimental values obtained from literature showed substantial variation, even for the same compound (Fig. 6d and e). This variance can be attributed to several factors. Simulations are typically based on idealized, defect-free models and do not account for real-world influences such as temperature, pressure, humidity, or the presence of guest molecules. Furthermore, experimental values can vary depending on synthesis routes, measurement techniques, and inconsistencies in reporting practices across different publications. These factors contribute to the broad range observed in the experimental data, in contrast to the uniformity of simulation outputs.

These findings highlight the importance of accounting for such discrepancies when integrating computational and experimental datasets in MOF research. As LLM-based text mining becomes more widely used for database construction, it will be critical to consider the contextual and methodological variability inherent in experimental data to ensure robust comparison and integration with simulation results.

4. Conclusion and perspectives

The integration of text mining into MOF research represents a paradigm shift in material design, synthesis, and optimization. This feature article highlights key advancements in text mining across the MOF landscape, spanning from traditional rule-based extraction to state-of-the-art LLM-based text mining. Conventional NLP and ML techniques, such as part-of-speech tagging and NER, have revolutionized data extraction by converting unstructured scientific literature into structured datasets, thereby uncovering critical synthesis trends. LLM-based text mining has further streamlined this process, significantly improving data accessibility and usability. By automating information retrieval, interpretation, extraction and curation, LLMs lower the technical barrier, making complex computational tools more accessible to researchers across various disciplines.

Despite these advancements, several challenges remain in fully harnessing text mining for MOF research. To address these challenges and further expand its capabilities, we propose four key directions for future advancements in text mining applications.

4.1 Integration of GUI in MOF informatics

Even prior to the release of ChatGPT, several pre-trained transformer models were available; however, their adoption was not very impactful. The emergence of ChatGPT (GPT-3.5) in 2022 marked a turning point, catalyzing the application of LLMs across diverse domains including computer science, materials science, biology, industry, and finance. While this rapid expansion can be partly attributed to technological advancements and improved model performance, a critical enabler was the development of a chat-style graphical user interface (GUI), which significantly lowered the barrier to entry

by allowing users to interact with LLMs through a simple, intuitive webpage. In this context, embedding a GUI with text-mined data offers a powerful approach to make structured materials datasets more accessible and interpretable to a broader research community.

Interactive GUI platforms such as the materials project[58] and the Cambridge Structural Database[59] have played a pivotal roles in recent advancements in AI-assisted materials informatics by enabling intuitive data retrieval. The GUI-based platforms are now being actively developed into other materials field, such as catalysts[60,61] and batteries[62] to support structure-performance visualization, and machine learning-assisted material screening. In the MOF domain, the QMOF database was integrated into Materials Project, providing DFT-derived properties (*e.g.*, optimized structures, bandgaps, and band structures).[63] Similarly, the recent update of the CoRE MOF 2025 database[64] introduced a streamlined web interface that allows users to simply drag and drop CIF files to compute geometric descriptors and predict properties such as water and thermal stability. Beyond simple data retrieval, integrating extracted data into chatbot-based GUIs (like ChatGPT) can assist users in better understanding the data and generating research ideas.[56,65]

Therefore, developing accessible and user-friendly GUIs will become increasingly important in materials science to ensure that a wider range of researchers can utilize available tools and data. Integrating text-mined information into interactive GUIs is able to eliminate the need for rigid, formally structured queries, lowering the barrier for non-experts. As these tools evolve, they hold the potential to support intuitive data exploration and significantly accelerate materials discovery.

4.2 Text mining-driven multi-agent AI systems

MOF synthesis process is inherently complex, requiring precise control over numerous variables such as metal precursors, organic linkers, solvents, and reaction conditions. Due to the intricate nature of laboratory workflows, AI is often applied only to isolated stages of research, leading to a disjointed process that heavily depend on human intervention. Multi-agent AI systems trained on domain-specific text-mined knowledge offer a more integrated and robust solution.[66] By assigning specialized AI agents to tasks such as reaction condition prediction, structure–property correlation, and synthesis planning, researchers can develop an autonomous and precise system capable of continuously optimizing MOF synthesis strategies based on structured knowledge extracted from the literature. A recent study by Yaghi *et al.* demonstrates a team of seven distinct AI research assistants which was assembled to optimize the crystallinity of MOFs and covalent organic frameworks (COFs).[67] Each AI assistant specialized in specific research tasks, including literature review, laboratory operations, and data interpretation, working collaboratively to expedite the discovery of optimal synthesis conditions. To develop such multi-agent AI systems, training each agent on text-mined data from specialized fields is essential, ensuring that AI-driven decision-making is grounded in accurate, context-specific scientific knowledge. A well-established text mining architecture enables multiple AI agents to collaborate effectively, thereby reducing reliance on human oversight and accelerating advancements in MOF research.

4.3 Integrating text mining into autonomous laboratory

An autonomous laboratory (A-Lab) aims to integrate robotics hardware and AI-driven software to accelerate materials discovery with minimal human control.[68–70] These systems are designed to plan, execute, and analyse experiments in a closed-loop workflow, optimizing synthesis conditions and improving reproducibility. The ultimate goal of A-Lab is to transition from manual experimental design to fully automated material synthesis and characterization, where AI-driven decision-making guides every step. However, current A-Lab primarily focus on reaction execution, while data curation and interpretation remain major bottlenecks, as synthesis and characterization results are often unstructured and require expert analysis. Integrating real-time text mining into A-Lab enables the extraction of structured insights from both scientific literature and *in situ* experimental data, enhancing active learning through continual data feeding and refinement while optimizing user-desired material synthesis by dynamically improving AI-driven synthesis planning, reaction parameter selection, and characterization analysis. For example, retrieval-augmented generation (RAG) is a framework that enhances LLMs by dynamically retrieving and integrating external, up-to-date information into their generative process,[71,72] thereby enabling A-Lab to autonomously refine experimental decisions and adapt synthesis protocols with enhanced accuracy. Notably, integrating accurate text mining into RAG transforms unstructured experimental and literature data into structured, actionable insights, significantly improving data interpretation in the feedback loop of A-Lab. The integration of text mining allows A-Lab to continuously update their knowledge base, facilitating machine-to-human communication through natural language, thereby accelerating exploration of vast MOF chemical spaces.

4.4 Multi-modal LLM-enhanced data mining for comprehensive MOF data extraction

In scientific literature, data can be represented in various forms, including 1D text, 2D images, and 3D chemical files (*e.g.* XYZ, crystallographic information file; cif, protein data bank; pdb), each providing unique insights and information.[73,74] Conventional text mining techniques primarily focus on extracting information from unstructured text, often neglecting critical data embedded in tables, graphs and files. However, many essential aspects of MOF research, such as synthesis conditions, characterization results, and structure–property relationships, are commonly presented in graphical plots, reaction schemes, and even simulation videos. For instance, nitrogen adsorption isotherms used to determine BET surface area are typically presented as graphical plots; powder X-ray diffraction patterns appear as a combination of graphs and textual annotations; and gas diffusion behaviors within MOF pores are often illustrated through molecular dynamics trajectory videos. Multi-modal LLMs, such as

Kosmos-2,[75,76] Flamingo,[77] LLaVA,[78] and PaLM-E,[79] have demonstrated the capability to process and interpret diverse data types, making them ideal for end-to-end MOF data mining. Data mining using multi-modal LLMs will enhance the feasibility and usability of extracted data by enabling seamless integration of information across different formats within scientific literature. This not only improves accessibility for experts conducting advanced analyses but also streamlines data interpretation for non-experts, bridging the gap between computational tools and experimental research.

Author contributions

Suyeon Bae: conceptualization, data curation, methodology, investigation, visualization, writing – original draft, writing – review & editing. Mingyu Jeon: data curation, methodology, investigation, writing – original draft, writing – review & editing, supervision. Hoi Ri Moon: conceptualization, writing – review & editing, supervision, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A2C3008908, RS-2025-02982993, and NRF-2021R1A6A10039823) and the Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (2020R1A6 C101B194).

References

1 H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, 1230444.
2 H.-C. Zhou, J. R. Long and O. M. Yaghi, *Chem. Rev.*, 2012, **112**, 673–674.
3 V. F. Yusuf, N. I. Malek and S. K. Kailasa, *ACS Omega*, 2024, **9**, 29947–29950.
4 M. Jeon, J.-S. Lee, M. Kim, J.-W. Seo, H. Kim, H. R. Moon, S.-J. Choi and J. Kim, *ACS Appl. Mater. Interfaces*, 2024, **16**, 62382–62391.
5 M. Kim, M. Pander and H. R. Moon, *ACS Appl. Electron. Mater.*, 2024, **6**, 3024–3038.
6 W. Xu and O. M. Yaghi, *ACS Cent. Sci.*, 2020, **6**, 1348–1354.
7 G. Hai and H. Wang, *Coord. Chem. Rev.*, 2022, **469**, 214670.
8 P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, *Chem. Mater.*, 2017, **29**, 2618–2625.
9 J. Park, H. Kim, Y. Kang, Y. Lim and J. Kim, *JACS Au*, 2024, **4**, 3727–3743.
10 K. Neikha and A. Puzari, *Langmuir*, 2024, **40**, 21957–21975.
11 J. Lee, W. Lee and J. Kim, *ACS Appl. Mater. Interfaces*, 2024, **16**, 723–730.
12 K. T. Mukaddem, E. J. Beard, B. Yildirim and J. M. Cole, *J. Chem. Inf. Model.*, 2019, **60**, 2492–2509.
13 G. H. Yi, J. Choi, H. Song, O. Miano, J. Choi, K. Bang, B. Lee, S. S. Sohn, D. Buttler and A. Hiszpanski, *Adv. Sci.*, 2025, 2408221.
14 X. Zhang, K. M. Jablonka and B. Smit, *Digital Discovery*, 2024, **3**, 1410–1420.
15 M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal and A. Valencia, *Chem. Rev.*, 2017, **117**, 7673–7761.
16 L. Ghadbeigi, J. K. Harada, B. R. Lettiere and T. D. Sparks, *Energy Environ. Sci.*, 2015, **8**, 1640–1650.
17 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
18 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. Jain, *Nature*, 2019, **571**, 95–98.
19 Q. Le and T. Mikolov, *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
20 T. Mikolov, K. Chen, G. Corrado and J. Dean, *arXiv*, 2013, preprint, arXiv:1301.3781, **https://arxiv.org/abs/1301.3781v3**.
21 L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder and A. Jain, *J. Chem. Inf. Model.*, 2019, **59**, 3692–3702.
22 J. Choi, K. Bang, S. Jang, J. Choi, J. Ordonez, D. Buttler, A. Hiszpanski, T. Yong-Jin Han, S. S. Sohn, B. Lee, K.-R. Lee, S. S. Han and D. Kim, *J. Mater. Chem. A*, 2023, **11**, 17628–17643.
23 H. Park, Y. Kang, W. Choe and J. Kim, *J. Chem. Inf. Model.*, 2022, **62**, 1190–1198.
24 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 5998–6008.
25 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
26 A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder and A. Jain, *Patterns*, 2022, **3**(4), 100488.
27 I. Beltagy, K. Lo and A. Cohan, *Proc. EMNLP-IJCNLP*, 2019, pp. 3615–3620.
28 S. Huang and J. M. Cole, *J. Chem. Inf. Model.*, 2022, **62**, 6365–6377.
29 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chem. Soc. Rev.*, 2025, **54**, 1125–1150.
30 V. T. da Silva, A. Rademaker, K. Lionti, R. Giro, G. Lima, S. Fiorini, M. Archanjo, B. W. Carvalho, R. Neumann and A. Souza, *arXiv*, 2024, preprint, arXiv:2411.03484, **https://arxiv.org/abs/2411.03484v1**.
31 Z. Zheng, N. Rampal, T. J. Inizan, C. Borgs, J. T. Chayes and O. M. Yaghi, *Nat. Rev. Mater.*, 2025, **10**, 369–381.
32 A. D. White, *Nat. Rev. Chem.*, 2023, **7**, 457–458.
33 J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman and S. Anadkat, *arXiv*, 2023, preprint, arXiv:2303.08774, **https://arxiv.org/abs/2303.08774v2**.
34 G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan and S. Wang, *arXiv*, 2024, preprint, arXiv:2403.05530, **https://arxiv.org/abs/2403.05530v5**.
35 A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang and A. Fan, *et al.*, *arXiv*, 2024, preprint, arXiv:2407.21783, **https://arxiv.org/abs/2407.21783v3**.
36 M. Livne, Z. Miftahutdinov, E. Tutubalina, M. Kuznetsov, D. Polykovskiy, A. Brundyn, A. Jhunjhunwala, A. Costa, A. Aliper, A. Aspuru-Guzik and A. Zhavoronkov, *Chem. Sci.*, 2024, **15**, 8380–8389.
37 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, *Nat. Mach. Intell.*, 2024, **6**, 161–169.
38 L. Chen, W. Wang, Z. Bai, P. Xu, Y. Fang, J. Fang, W. Wu, L. Zhou, R. Zhang and Y. Xia, *arXiv*, 2024, preprint, arXiv:2406.18045, **https://arxiv.org/abs/2406.18045v3**.
39 A. Li, R. Bueno-Perez and D. Fairen-Jimenez, in *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*, ed. G. Sastre and F. Daeyaert, Wiley, 2023, ch. 8, pp. 201–232.
40 S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit and J. Kim, *J. Chem. Inf. Model.*, 2018, **58**, 244–251.
41 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chem. Mater.*, 2014, **26**, 6185–6192.

42 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling and J. S. Camp, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
43 A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner and H. J. Kulik, *Sci. Data*, 2022, **9**, 74.
44 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
45 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
46 L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S. M. Moosavi, J. L. Cordiner, J. C. Cole and P. Z. Moghadam, *Chem. Mater.*, 2023, **35**, 4510–4524.
47 Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich and M. Tsotsalas, *Angew. Chem., Int. Ed.*, 2022, **61**, e202200242.
48 D. Lee, H. Mizuseki, J. Choi and B. Lee, *Commun. Mater.*, 2025, **6**, 1–13.
49 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, *Nat. Commun.*, 2024, **15**, 1418.
50 M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed and M. A. Sattar, *Sci. Rep.*, 2025, **15**, 13755.
51 Y. Huang, K. Tang, M. Chen and B. Wang, *arXiv*, 2024, preprint, arXiv:2404.15777, **https://arxiv.org/abs/2404.15777v4**.
52 C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, T. Gao, E. Li, K. Tang, Z. Cao, T. Zhou, A. Liu, X. Yan, S. Mei, J. Cao, Z. Wang and C. Zheng, A Survey on Multimodal Large Language Models for Autonomous Driving, *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, 958–979.
53 S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu and Q. Wen, *arXiv*, 2024, preprint, arXiv:2403.18105, **https://arxiv.org/abs/2403.18105v2**.
54 Q. Ren, Z. Jiang, J. Cao, S. Li, C. Li, Y. Liu, S. Huo, T. He and Y. Chen, *arXiv*, 2024, preprint, arXiv:2405.13025, **https://arxiv.org/abs/2405.13025v2**.
55 J. Lee, N. Stevens and S. C. Han, *Neural Comput. Appl.*, 2025, 1–15.
56 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
57 Y. Kang, W. Lee, T. Bae, S. Han, H. Jang and J. Kim, *J. Am. Chem. Soc.*, 2025, **147**, 3943–3958.
58 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
59 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
60 J. Fujima, Y. Tanaka, I. Miyazato, L. Takahashi and K. Takahashi, *React. Chem. Eng.*, 2020, **5**, 903–911.
61 M. Kuwahara, J. Fujima, K. Takahashi and L. Takahashi, *Digital Discovery*, 2023, **2**, 775–780.
62 S. Huang and J. M. Cole, *Sci. Data*, 2020, **7**, 260.
63 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.
64 G. Zhao, L. M. Brabson, S. Chheda, J. Huang, H. Kim, K. Liu, K. Mochida, T. D. Pham, G. G. Terrones and S. Yoon, *Matter*, 2025, **8**(6), 102140.
65 Y. Kang and J. Kim, *Nat. Commun.*, 2024, **15**, 4705.
66 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, **6**, 525–535.
67 Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes and O. M. Yaghi, *ACS Cent. Sci.*, 2023, **9**, 2161–2170.
68 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, *Chem. Rev.*, 2024, **124**, 9633–9732.
69 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, *Nature*, 2023, **624**, 86–91.
70 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
71 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih and T. Rocktäschel, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 9459–9474.
72 J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques and A. D. White, *arXiv*, 2023, preprint, arXiv:2312.07559, **https://arxiv.org/abs/2312.07559v2**.
73 M. C. Ramos, C. J. Collison and A. D. White, *Chem. Sci.*, 2025, **16**, 2514–2572.
74 D. Flam-Shepherd and A. Aspuru-Guzik, *arXiv*, 2023, preprint, arXiv:2305.05708, **https://arxiv.org/abs/2305.05708v1**.
75 S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed and B. Patra, *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 72096–72109.
76 Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma and F. Wei, *arXiv*, 2023, preprint, arXiv:2306.14824, **https://arxiv.org/abs/2306.14824v3**.
77 J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican and M. Reynolds, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 23716–23736.
78 H. Liu, C. Li, Q. Wu and Y. J. Lee, *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 34892–34916.
79 D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu and W. Huang, *arXiv*, 2023, preprint, arXiv:2303.03378, **https://arxiv.org/abs/2303.03378v1**.

